CHARACTERIZING AND DETECTING UNREVEALED ELEMENTS OF NETWORK SYSTEMS

DISSERTATION

James A. Leinart, Lieutenant Colonel, USAF

AFIT/DS/ENS/08-01W

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views expressed in this dissertation are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense or the United States Government.

AFIT/DS/ENS/08-01W

CHARACTERIZING AND DETECTING UNREVEALED ELEMENTS OF NETWORK SYSTEMS

DISSERTATION

Presented to the Faculty
Department of Operational Sciences
Graduate School of Engineering and Management
Air Force Institute of Technology
Air University
Air Education and Training Command
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy in Operations Research

James A. Leinart, B.S., M.S.
Lieutenant Colonel, USAF

March 2009

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AFIT/DS/ENS/08-01W

CHARACTERIZING AND DETECTING UNREVEALED ELEMENTS OF NETWORK SYSTEMS

James A. Leinart, B.S., M.S.
Lieutenant Colonel, USAF

Approved:

Dr. Richard F. Deckro Date
Committee Chairman

Dr. James T. Moore Date
Committee Member

Lt Col David R. Denhard Date
Committee Member

Dr. Marcus B. Perry Date
Committee Member

Dr. Robert F. Mills Date
Committee Member

M.U. Thomas
Dean

AFIT/DS/ENS/08-01W

Abstract

This dissertation addresses the problem of discovering and characterizing unknown elements in network systems. Klir (1985) provides a general definition of a system as "... a set of some *things* and a *relation* among the things" (p. 4). A system, where the 'things', i.e. nodes, are related through links is a network system (Klir, 1985). The nodes can represent a range of entities such as machines or people (Pearl, 2001; Wasserman & Faust, 1994). Likewise, links can represent abstract relationships such as causal influence or more visible ties such as roads (Pearl, 1988, pp. 50-51; Wasserman & Faust, 1994; Winston, 1994, p. 394).

It is not uncommon to have incomplete knowledge of network systems due to either passive circumstances, e.g. limited resources to observe a network, active circumstances, e.g. intentional acts of concealment, or some combination of active and passive influences (McCormick & Owen, 2000, p. 175; National Research Council, 2005, pp. 7, 11). This research provides statistical and graph theoretic approaches for such situations, including those in which nodes are causally related (Geiger & Pearl, 1990, pp. 3, 10; Glymour, Scheines, Spirtes, & Kelly, 1987, pp. 75-86, 178-183; Murphy, 1998; Verma & Pearl, 1991, pp. 257, 260, 264-265). A related aspect of this research is accuracy assessment. It is possible an analyst could fail to detect a network element, or be aware of network elements, but incorrectly conclude the associated network system structure (Borgatti, Carley, & Krackhardt, 2006). The possibilities require assessment of the accuracy of the observed and conjectured network systems, and this research provides a means to do so (Cavallo & Klir, 1979, p. 143; Kelly, 1957, p. 968).

Acknowledgements

A significant portion of this effort was accomplished in Ohio; how fitting that it should bear out the state motto, "With God, All Things Are Possible." I humbly acknowledge God's grace, love and sustainment throughout this process. Much of His provision came through the following people:

My wife, the love of my life, who walked with me through this journey, and marvelously cared for our children.

My children who gave me love, invoked my compassion and modeled joyful living.

My parents who supported and loved me before, during and after the doctoral program.

My extended family who supported and encouraged me.

My friends, AFIT classmates and the HQ USAF/A9 team who also supported and encouraged me.

My advisor, Dr. Deckro, who guided me along this road and cared for me beyond the academic realm.

Lt Col Denhard, though called to serve elsewhere, provided faithful support and counsel to me.

My committee, Dr. Moore, Dr. Perry and Dr. Mills (and the Dean's representative, Dr. Baker) who made this dissertation a reality.

Students and faculty from AFIT and other universities who took time to answer my questions, provide insight and encourage me.

Thank you.

James A. Leinart

Table of Contents

Page

Page

Page

List of Figures

List of Tables

CHARACTERIZING AND DETECTING
UNREVEALED ELEMENTS OF NETWORK SYSTEMS

1. Introduction

1.1 Background

Network representations of real world entities and processes abound. Network nodes may represent entities with the links describing some type of relationship, e.g. causal, between the nodes. Information about network membership and structure can be partial and varied depending on both the network type and the network observation process. Consequently, the resulting uncertainties may lead to risks when courses of action are planned and implemented.

Network representations can also be used to model a complex environment or system each composed of systems. These network systems can be, among other types, social or computational in nature (Ferrand, Mounier & Degenne, 1999; Klir, 1985). Given the complexity and size of some network systems, it may be difficult to fully characterize them and understand their workings (National Research Council, 2005, pp. 7, 11). Consider, for example, the covert and dynamic (due in part to accession and attrition) al-Qa`ida terrorist network. It was not until five years after 9/11 that `Atiyah, a high-ranking official in the al-Qa`ida terrorist network, was revealed, at least within the open press (Combating Terrorism Center, 2006; McCormick & Owen, 2000, pp. 186-187). Information of this type is extremely valuable in effectively combating terrorism and/or properly modeling related social networks; hence, an approach to detect such unknown key actors and their ties would be of great assistance in developing and choosing courses of actions against such networks.

1.2 Problem Statement

From a network analysis perspective, the concepts above yield an interesting problem area: Detecting and characterizing unrevealed elements in network systems, where some nodes may affect others, and network information may be partial. Such is the crux of this research.

Researchers from a variety of fields have examined hidden nodes and links in different contexts. Psychologists have endeavored to determine how individuals infer hidden causes (Kushnir, Gopnik, Schulz, & Danks, 2003). Individuals in the artificial intelligence (AI) community have pursued identifying previously unaccounted mechanisms (e.g. Doyle, 1989). The AI community relies heavily upon Bayesian networks, and one of the related problems is the presence of hidden, i.e. unobserved, variables (Binder, Koller, Russell, & Kanazawa 1997). Social network analysts have also researched the impact of missing links and nodes; additionally, some preliminary work has been accomplished in statistical modeling of social processes where nonrespondents (i.e. hidden actors) are present (Borgatti *et al.*, 2006; Robins, Pattison & Woolcock, 2004).

Additionally, there is an area in systems science known as reconstructability analysis (RA), which refers to the process of examining the possibilities of reconstructing desirable properties of overall systems using knowledge of respective properties of their various subsystems (Cavallo & Klir, 1979, p. 143). RA addresses two problems: identification and reconstruction. The identification problem attempts to infer an overall, unknown system from its subsystems, while the reconstruction problem aims at determining subsytems that can adequately reconstruct a known, higher-level system (Klir, 1985, p. 212). If a system is modeled as a graph G with n nodes, then one type of subsystem could be a vertex-deleted subgraph of G, which is a subgraph containing $n-1$ nodes (and their incident links) of G (Bondy & Hemminger, 1977, p. 227; Kelly, 1957, p. 961). Given this representation, graph theory contains a form of the identification problem known as the reconstruction conjecture.

The conjecture asserts that any two graphs, G_1 and G_2 (each with $n > 2$ nodes), having the same (unlabeled) vertex-deleted subgraphs are isomorphic, i.e. the graph is reconstructable (Kelly, 1957, p. 968). Consequently, these concepts are appropriate for characterizing the relation of nodes in network systems, so the network structure can be determined. The literature review, in the next chapter, details related efforts to address hidden and unknown entities.

What emerges from these, and other, research efforts are viable techniques for addressing portions of the unrevealed elements problem. This research effort attempts to synthesize previous contributions and create novel detection means in order to provide a macroscopic approach relevant to network systems, to include those viewed via an observation process that may capture only partial network information.

1.3 Contributions

Given the above problem area, the contributions of this dissertation include:

1. A method to identify and detect nodes in a social network where influence among nodes is not explicitly considered. The network examined contains nodes connected by only a single link, and there are no links from a node to itself. Consequently, the network may be represented as a simple graph (West, 2001, p. 2). Furthermore, a static network structure is assumed.
2. A paradigm for incorporating causal, stochastic and temporal aspects within the context of network reconstruction.
3. A process to identify possible structures of a social influence network and to detect potential unrevealed individuals. A social influence network will be represented by a directed graph.

1.4 Conclusion

Network systems can be complex and large. Adding to such characteristics the notions of secrecy and partial observability yields a rather daunting problem for determining structure and behavior. This research aims at uncovering the structure of such network systems to provide more insight into the membership and structure of such systems. This gained knowledge can then aid course of action creation, selection and risk quantification.

The remainder of this document is as follows: Chapter 2 discusses literature germane to the problem area, showing the gaps which this effort's contributions intend to fill; Chapter 3 provides methodological assumptions and descriptions; Chapters 4-6 present results associated with each contribution; and Chapter 7 summarizes the dissertation and provides areas for future research.

2. Review of the Literature

As introduced in Chapter 1, there are many different network types; consequently, addressing unrevealed elements requires a diversity of methods depending upon the network under evaluation. Many network types are composed of entities that can (or appear to) affect each other, i.e. network topology represents correlation among nodes. Hence incorporation of data analysis techniques, in addition to structural analysis, may be necessary to reveal network elements. Additionally, given the adage that correlation is not the same as causation, it is beneficial to examine causality concepts and how they can be leveraged to address the unrevealed element problem. This chapter reviews related concepts and previous efforts in these areas.

2.1 Communication and Social Networks

The concepts in this dissertation are germane to many network types, but the primary application is a social network. A social network consists of actors and their associated ties. The actors can be individuals or larger social units (e.g. collectives), and ties are linkages between two actors. Furthermore, the entire set of a specific kind of tie for the actors in a social network is referred to as a relation (Wasserman & Faust, 1994, pp. 9, 17-18, 20). Communication networks are essentially a type of social network with individuals as the actors, and the relation is defined by their set of communication ties (Rogers & Kincaid, 1981, pp. 94, 346; Wasserman & Faust, 1994, pp. 18, 20). Communication networks often exhibit some form of structure and, "Cliques are the most important single aspect of communication structure in a system" (Rogers & Kincaid, 1981, p. 146). Furthermore, a system in a communication network, "... is a set of interrelated parts coordinated to accomplish a set of goals" (Rogers & Kincaid, 1981, p. 348). In this social science context, cliques are subsystems whose individuals communicate with each other more often than with other individuals in the system. Such cliques are not necessarily the same as the graph theoretic concept of cliques, in which all vertices are connected; rather the notion of a $k-$plex may be appropriate, i.e. every member of a subsystem, containing n members, communicates with at least $n - k$ other members. Consequently, the traditional graph theoretic clique corresponds to a 1-plex, i.e. $k = 1$ (Seidman & Foster, 1978, pp. 142-143). Two roles related to cliques are bridges and liaisons, who are individuals that connect cliques; however, according to Rogers and Kincaid (1981), the former are members of one of the cliques they connect, whereas the latter are not members of either clique (Rogers & Kincaid, 1981, pp. 146, 346-347). Ross and Harary (1955) discuss the fact (mentioned in other authors' efforts) that liaisons can be graphically represented as articulation points, i.e. cut vertices (Harary & Norman, 1953, p. 27; Ross & Harary, 1955, p. 253; Weiss & Jacobson, 1955, p. 664; West, 2001, pp. 23, 575). Terrorist organizations can be considered a form of communication or social networks. As indicated by Carley, Dombroski, Tsvetovat, Reminga, and Kamneva (2003) and Krebs (2002), such networks are cellular and distributed. The cells can be represented as groups or cliques with liaisons or bridges linking them together.

In the context of communication networks, heterophily and homophily refer to the degree of interaction among individuals with different or similar attributes, respectively (Rogers & Kincaid, 1981, pp. 346-347). Liu and Duff (1972) provided some empirical evidence that information diffusion through a network requires a degree of heterophily among certain members, stating that "It is through the *infrequent*, but *strategic*, contacts...that such information gains remarkably wide circulation" (Liu & Duff, 1972, p. 366). Granovetter (1973) also noted that diffusion is dependent upon some amount of weak ties among the communication network members, with the strength of a tie defined as some combination of its duration, emotional intensity, intimacy and reciprocal services (Granovetter, 1973, pp. 1361, 1366). Consequently, Rogers and Kincaid (1981) stated that weak ties involve bridges and liaisons (Rogers & Kincaid, 1981, p. 128). From the literature presented in this and other sections, revealing bridges is important since they play a key role in networks.

The literature indicates characteristics that one might expect a liaison (and by assumption, a bridge) to exhibit, i.e in regression terminology, variables that explain a liaison (Dillon & Goldstein, 1984). One of the first explanatory concepts examined is the level of credibility of an individual. Hovland, Janis, and Kelly (1953) asserted that credibility of a communicator depends upon perceived expertness and trustworthiness (Hovland, Janis, & Kelly, 1953, p. 21). Berlo, Lemert, and Mertz (1969) performed factor analysis with respect to a similar notion, i.e. "evaluating message sources" (Berlo, Lemert, & Mertz, 1969, p. 565). Two resulting factors were safety and qualification; these are somewhat compatible with the concepts of expertise and trustworthiness (Berlo *et al.*, 1969, p. 574). Rogers and Bhowmik (1971) build upon these concepts in their statement that a source viewed as having qualification credibility is often heterophilous with respect to the receivers, while a source perceived as having safety credibility is frequently highly homophilous to the receivers (Rogers & Bhowmik, 1971, p. 534). Rogers and Bhowmik (1971) claim that multiple studies, summarized in Rogers and Shoemaker (1970), support the logic of their above statement. While it appears intuitive that an individual between two groups should be perceived as credible, such a concept may not entirely enable distinguishing a liaison from a nonliaison, as shown in a study by Schwartz and Jacobson (1977). Thus credibility may be a probable rather than a sufficient condition for an individual to be a bridge; furthermore, credibility is only one of multiple explanatory concepts.

The expertness aspect of credibility can be represented by an education level and experience level components, which are similar to the concepts of knowledge and experience required by integrators in an organization (Lawrence & Lorsch, 1967, pp. 146-147). Intuitively, a higher education level is associated with greater expertise. Schwartz and Jacobson (1977) indicated there was no significant difference between liaisons and non-liaisons regarding education level; however, the study was performed in a university college among faculty and academic administrators with a rank of instructor or higher. Such a sample appears to create some bias toward the ineffectiveness of a statistic based on education level. An unpublished research report by McPhee and Meyersohn (1951) indicated that emerging opinion leaders were young people with some education who were mobile but kept in contact with their families and illiterate neighbors (as cited in Katz & Lazarsfeld, 1964, pp. 127-128). The study was conducted in Lebanon among rural peasants who could not read. Despite the small sample size and the narrow cultural context, it appears plausible that credibility, as associated with an individual's influence, e.g. Hovland and Weiss (1951), may be impacted by education level. Likewise, it is not unreasonable to assume that the longer an individual has been in an organization performing various tasks, i.e. the more experience the individual has, the higher the credibility of the individual (Hovland *et al.*, 1953, pp. 21-22). This correlation may not always hold as discussed in Lawrence and Lorsch (1967); nevertheless, the association appears plausible in certain situations.

Determining trustworthiness in social networks can be difficult since the statements and subsequent actions of individuals are not always consistently observable. Consequently, for some types of research, the explanatory concept of experience may need to represent both the expertness and trustworthiness of an individual.

As noted above, credibility and influence are related; furthermore, a gatekeeper position may be considered a bridge position as defined in this research (Conway, 1997, p. 228). Additionally, based on various studies (e.g. Eisenstadt (1952) and Horsfall and Arensberg (1949)), Katz and Lazarsfeld (1964) noted that a gatekeeper may exert influence within a group (Katz & Lazarsfeld, 1964, p. 123). While the accuracy of such a conclusion (based on the studies) may be debatable, this research assumes that influence related explanatory variables appear plausible for examining bridges. The influence related variables examined in this research are oriented toward leadership and status.

Katz and Lazarsfeld (1964) defined opinion leaders as individuals who transmitted influences of an interpersonal nature in common situations regarding topics such as fashions or public affairs (Katz & Lazarsfeld, 1964, p. 219). Specifically, Katz and Lazarsfeld (1964) examined influence among women; however, the derived results can be generalized. The factors that Katz and Lazarsfeld (1964) examined were life-cycle, social and economic status, and gregariousness. Life-cycle corresponds to the phase of an individual's life, with age a component of the factor (Katz & Lazarsfeld, 1964, p. 221). A liaison might be expected to be older than a non-liaison, e.g. Eisenstadt (1952); however, at least one study found no significance in age between the two (Schwartz & Jacobson, 1977, pp. 161, 168). Alternatively, it might be reasonable to assume that a liaison is usually younger than a non-liaison (McPhee & Meyersohn (1951) as cited in Katz & Lazarsfeld, 1964, pp. 127-128). Despite these mixed results, age may be important in a bridge context.

Status can be considered an influence related explanatory variable and can be decomposed into various aspects to include education (e.g. Katz & Lazarsfeld, 1964, p. 226). Rank can be viewed as a measure of status; MacDonald (1976) found that, on average, liaisons were higher in grade than the non-liaisons with whom they had contact (MacDonald, 1976, p. 372). Lasswell and Kaplan (1950) stated, "power is a type of influence" (p. 84); furthermore, they assert group leaders hold power (Lasswell & Kaplan, 1950, p. 152). Consequently, leadership (and arguably supervision) is another aspect of influence, and it is natural to assume that bridges could be leaders and supervisors (Eisenstadt, 1952; Horsfall & Arensberg, 1949, pp. 24-25; MacDonald, 1976, p. 372).

Katz and Lazarsfeld (1964) examined the impact of gregariousness, i.e. the number of contacts an individual has, as it related to opinion leadership, and found some evidence that as gregariousness increased so did opinion leadership (Katz &

Lazarsfeld, 1964, pp. 223, 243, 259, 288). Their research motivates consideration of gregariousness when characterizing bridges. Gregariousness is an attribute that, when expressed in terms of a social network, can be approximated by the degree of a node, i.e. the number of (undirected) links for an actor. In order to compare node degree across groups of different sizes, the degree $deg(v_i)$ of a node v_i in a group containing n nodes is determined by the following index $\frac{deg(v_i)}{(n-1)}$ (Proctor & Loomis, 1951, pp. 570-571; Wasserman & Faust, 1994, pp. 178-179).

Since bridges can be viewed as gatekeepers, another influence related explanatory variable is the amount to which a network node depends upon another node, i.e. a gatekeeper, in order to communicate with other network nodes (Freeman, 1980, p. 587). Freeman (1980) developed a measure to capture this pair-dependency; furthermore, both betweenness and closeness centrality measures can be derived from the pair-dependency measure. This research focuses on identifying unrevealed bridges where their links to other groups is unknown; however, it is plausible to assume non-liaison nodes might depend upon the liaison node for intra-group communication as well as inter-group communication. A standardized version of the betweenness index of Freeman (1977) is appropriate in such situations. Specifically, for a group of n nodes, with g_{ik} representing the number of shortest paths between v_i and v_k, and $g_{ik}(v_j)$ denoting the number of shortest paths between v_i and v_k that contain v_j, the index is $\frac{\sum_{k=1}^{n} \frac{g_{ik}(v_j)}{g_{ik}}}{\frac{(n-1)(n-2)}{2}}$, $i \neq j \neq k$ and $i < k$. (Freeman, 1977, pp. 37-38; Wasserman & Faust, 1994, p. 190).

It is also plausible to consider an explanatory variable representing the similarity of a group member to every other group member with respect to a combination of the previously given explanatory variables. Such a variable is in keeping with the previously discussed concepts of homophily and heterophily (Rogers & Bhowmik, 1971).

While there exists literature devoted to identifying liaisons and their attributes and functionality, the existence of liaisons is somewhat presupposed (MacDonald, 1976; Ross & Harary, 1955; Schwartz & Jacobson, 1977). For example, MacDonald (1976) and Schwartz and Jacobson (1977) used a matrix manipulation method reported in Weiss and Jacobson (1955) to identify network groups. Once the groups were identified, the individuals who connected them could be ascertained. These connection individuals were then compared, on the basis of various attributes, to individuals in the network who were not liaisons. In some instances, there emerge concepts that appear to be applicable not only to social networks, but also other network types, regarding the distinction of liaisons. The concepts include location, level of interaction, transferability (i.e. heterogeneity), status (due perhaps to nomination or certification) and character (e.g. gregariousness) (Freeman, 1980; Granovetter, 1973; Katz & Lazarsfeld, 1955, pp. 115, 118, 127, 220-228; Lewin, 1952, pp. 461-462; Liu & Duff, 1972; MacDonald, 1976, pp. 370-372; Rogers & Kincaid, 1981, pp. 30, 128, 146, 346-347; Schwartz & Jacobson, 1977, pp. 166, 169). An example of a liaison in a non-social network of networks is a router that links computer networks that are not similar (Englander, 2003).

The relevant literature does not appear to directly address the following problem: Given a network of groups, for which the information about each group consists of the individuals composing the group, attributes of the individuals, and (only) intra-group relations; identify which individual, possibly more than one, is a bridge. Additionally, for the same given information, the literature does not seem to contain the approach of this effort to detect that a bridge is missing from a group, i.e. infer the existence of an unknown bridge from group data that does not contain the bridge in its membership, the bridge's attributes, nor the bridge's contacts.

2.1.1 Influence in Networks

Another aspect to communication networks is the notion of influence. As noted by Liu and Duff, there are two parts to determining the effectiveness of communication: how far and quickly information diffuses, and the extent to which attitudes and behavior are changed (Liu & Duff, 1972, p. 365). Rogers and Bhowmik (1971) proposed that effective communication is maximized when there is the proper mix of heterophily and homophily for relevant variables between the source and receiver. Furthermore, they posit that *status inconsistent* individuals may be apt at fulfilling the liaison role in a social structure, e.g. such actors may be suited to link heterophilous cliques (Lenski, 1954; Rogers & Bhowmik, 1971, pp. 532-533). Rogers and Kincaid (1981) summarize, based on previous works (e.g. Granovetter (1973), Liu & Duff (1972) and Epstein (1961)) that strong links, i.e. within homophilous groups, are apt for influence; while, weak links, i.e. between heterophilous individuals/groups, permit new information (innovation) diffusion. However, Rogers and Kincaid (1981) noted that the strength of weak ties does not hold for every culture; therefore, any proposed behavioral science theories need to be validated across cultural boundaries (Rogers & Kincaid, 1981, pp. 243-247).

Merton (1957) derived the following notion for interpersonal influence from concepts of power presented by Goldhamer and Shils (1939).

> Interpersonal influence refers to the direct interaction of persons in so far as this affects the *future* behavior or attitude of participants (such that this differs from what it would have been in the absence of interaction) (Merton, 1957, p. 415).

Concerning the relationship between causality and influence, Merton (1957) notes the analysis of influence by March (1955) who, in turn, draws from concepts by Simon (1952, 1953). March (1955) argues that an influence relation set is a proper subset of a causal relation set, and both sets involve asymmetrical relations and an ordering of the relations. Simon highlights the notion of asymmetry between the influencer and influencee, when he implements the definition of 'influence process' given by Lasswell and Kaplan (1950) (Lasswell & Kaplan, 1950, pp. 71; March, 1955, pp. 436-437; Merton, 1957, pp. 415-416; Simon, 1952, pp. 517, 520; Simon, 1953, pp. 503-504). Furthermore, there are similarities between influence diagrams and causal networks, e.g. both are represented by directed acyclic graphs (DAGs), contain chance nodes and deal with conditional independencies (Geiger & Pearl, 1990, p. 3; Pearl, Geiger, & Verma, 1990, pp. 67-68). Consequently, in certain social network contexts, it appears reasonable to address influence via causal concepts and techniques (addressed in another section of this chapter).

Social influence in networks has been examined for many years. French (1956) examined the impact of the structure of a group's interpersonal relations (modeled as a digraph) on the group's process of influence. From this theory and the work of others, Marsden and Friedkin (1994) denoted the influence model as $\mathbf{y}_{t+1} = \mathbf{W}\mathbf{y}_t$; where $\mathbf{y}_t$ represents the attitudes of individuals at time t and $\mathbf{W}$ corresponds to coefficients of influence (Marsden & Friedkin, 1994, p. 10). Friedkin (1990) referred to the generalized version, $\mathbf{y} = \alpha\mathbf{W}\mathbf{y} + \beta\mathbf{X}\mathbf{b} + \mathbf{u}$, as the network model (Friedkin, 1990, p. 317). Friedkin (1990) noted that the model was provided in earlier literature, and Marsden and Friedkin (1994) comment on the model's appeal since it addresses influence from both endogenous, i.e. network effects, and exogenous perspectives via, the first and second term, respectively, of the right hand side of the equation. Path analysis and structural equation models (discussed in another section of this chapter) have also been used to address causality and social influence, e.g. Duncan, Haller and Portes (1968). Despite the discussion of causal principles in social network analysis (e.g. Doreian, 2001; Haller and Butterworth, 1960), the literature does not appear to contain the application of a causal analysis method to derive social influence network structures, at the interpersonal level, or detect hidden individuals in such networks.

2.2 Incomplete Information in Networks

Information about a social network's topology may be incomplete (e.g. hidden links or nodes) or contain uncertainties (e.g. does an extant link really exist, or a link exists but between which two nodes) (Butts, 2003; Robins *et al.*, 2004). Social network topological uncertainty can arise as a result of intentional and unintentional actions. An intentional action may be a consequence of a group's desire for secrecy, e.g. a terrorist network (McCormick & Owen, 2000). Alternatively, two individuals completing a survey may have different perceptions regarding network relationships (Banks & Carley, 1994; Killworth & Bernard, 1976; Krackhardt, 1987). The topic has been canvassed reasonably well in the social sciences literature, and the following review provides a few examples.

In the context of unintentional actions, Banks and Carley (1994), "...define[d] a probability measure for network-valued random variables" (p. 121). Consequently, when provided with a random sample of networks with some distribution (containing unknown parameters), one can obtain a maximum likelihood estimate of the 'commonly perceived' network, perform goodness-of-fit tests, test hypotheses and develop confidence regions (Banks & Carley, 1994, pp. 121-123, 128-131, 135-137). Interesting hypotheses may involve concepts such as subnetworks and uniqueness of the commonly perceived network (Banks & Carley, 1994, pp. 132-133).

In response to the situation where network data is incomplete, techniques have been developed to 'fill in the gaps'. Social networks often have incomplete data in the form of missing links. If the links between nodes have a strength attribute, then Burt (1987) commented that, "The implication is that the missing network data can be replaced with quantitative data indicating a weak relation" (p. 63). Butts (2003) has gone a step further by addressing not only missing links, but also extant links that are incorrect; all of which are caused by inaccurate informant reports (Butts, 2003, p. 110). Thus, his approach is concerned with determining the true underlying network, as was the approach of Banks & Carley (1994). The method Butts (2003) employed is relevant to situations involving uncertainties, e.g. intelligence regarding networks (Butts, 2003, p. 105).

In addition (or prior) to correcting an incomplete or inaccurate network topology, it would be useful to know the associated impact. Robins *et al.* (2004) developed models to assess the impact of missing data in determining various substructures of the network under investigation (Robins *et al.*, 2004, pp. 257, 272-275, 277-278).

The missing data includes nodes, i.e. non-respondents, and associated links (Robins *et al.*, 2004, pp. 264, 266, 277-278). The issue of which nodes to include in a network analysis is also discussed (Robins *et al.*, 2004, pp. 258-260). This issue is known as the boundary specification problem, which addresses inclusion of known individuals and their ties, rather than detection of unknown elements (Laumann, Marsden & Prensky, 1983). Consequently, Robins *et al.* (2004) do not address detecting missing nodes; however, they mention two situations where links could be added. The first is attributed to Stork and Richards (1992) and adds (i.e. reconstructs) a directed link from one node to another if the opposite direction link exists between the nodes (Stork & Richards, 1992, pp. 197-200). Second, if two nodes are connected to the same (non-responding) node, then adding a link between the two original nodes may be plausible (Robins *et al.*, 2004, pp. 260, 263).

Costenbader and Valente (2003) examined the stability of various centrality measures when the underlying social network data from which samples are obtained contain missing elements, e.g. non-respondents. Kossinets (2006) and Borgatti *et al.* (2006) performed missing data analysis that included not only edges, but also nodes and gave results that, for certain conditions, showed significant impact on estimates of network level statistics. Kossinets (2006) examined the impact of missing nodes on the following network statistics:

> mean vertex degree...; clustering...; assortativity...; fractional size of the largest connected component...; and average path length (mean geodesic distance) between all pairs of vertices in the largest component... (p. 254)

Borgatti *et al.* (2006) examined the impact of missing and additional data on centrality measures: degree, betweenness, closeness and eigenvector. The authors point out that centrality measures are robust when the data error is small, but context determines the acceptable level of robustness (Borgatti *et al.*, 2006, pp. 129, 134-135). This statement underscores the importance of this research effort, because a single undetected network node or edge (i.e. node or edge addition error) can have a critical impact on the ability to predict and manage associated risks.

Steinley and Wasserman (2006) examined the plausibility of identifying hidden links and nodes by assuming a generative model as appropriate for the application at hand, e.g. Bernoulli for a terrorist network, and then determining if network samples have statistics congruent with the conjectured distribution. The authors note that disagreement could indicate an incorrect generative model, or missing nodes and links, that if added, would yield network statistics in line with the generative model (Steinley & Wasserman, 2006, pp. 9-10). This dissertation follows in the vein of Steinley and Wasserman (2006); specifically, an approach that can address the presence of unknown nodes (and their incident links). The importance of such a technique is validated by the above research highlighting the impact of missing network elements.

2.3 Data Analysis

There are statistical analysis techniques able to address, in some form or fashion, issues germane to networks with unrevealed elements. Dillon & Goldstein (1984) noted that multivariate analysis includes a broad range of techniques designed to help analyze "simultaneous relationships among variables" (p. 2). The techniques can be decomposed into dependence and interdependence methods. Dependence methods involve explicating or predicting measures predicated on a set of predictor variables. Alternatively, interdependence methods are less predictive and attempt to show insights regarding the underlying structure of the data via simplification (e.g. data reduction). Dependence methods include: multiple regression, discriminant analysis and logistic regression. Interdependence methods include factor analysis, principal component analysis, multidimensional scaling and cluster analysis (Dillon & Goldstein, 1984, pp. 19-20).

2.3.1 Interdependence Methods

2.3.1.1 Factor Analysis

Factor analysis focuses on identifying structure from a set of observed variables. It includes a variety of techniques and in the broadest sense even principal component analysis can be considered a type of factor analysis. Broadly speaking, factor analysis provides three primary functions:

1. Reducing the number of variables while maintaining the largest possible amount of the original information, i.e. accounting for most of the variability.
2. Searching for distinct qualitative and quantitative characteristics in large amounts of data.
3. Testing hypotheses about such distinctions.

(Dillon & Goldstein, 1984, pp. 20, 23-24, 53, 56-57). It is important to note that factors are qualitative and thus cannot be observed (Dillon & Goldstein, 1984, pp. 53-54, 57, 60; Long, 1983, p. 11).

Principal component analysis attempts to determine the factor dimension of the data with respect to the total variance. The usual goal is to use the fewest possible principal components to account for the majority of the total variation. The components are linear combinations of the original variables. Additionally, the extracted components are orthogonal; consequently, they are uncorrelated, and this facilitates their use in other techniques such as regression (Dillon & Goldstein, 1984, pp. 8, 24-25, 27).

Factor analysis, by design, is aimed at uncovering unobservable (i.e. latent) variables by examining observed variable covariation. However, there are two perspectives from which to view the process of discovering unobservable variables. The first perspective is to consider the unobservable factors as a perfect function of observable variables (i.e. no measurement error), as in principal components analysis.

The second viewpoint is treating the observable indicators, each having an error term, as a function of the latent variable(s), which is the case in common factor analysis. The difference in perspectives is subtle, but important. Factor analysis can also be decomposed into both exploratory and confirmatory uses. Exploratory factor analysis is accomplished when an analyst is searching for an underlying structure to the data without an *a priori* theoretical hypothesis. Confirmatory factor analysis is conducted when an analyst has prior theoretical information on the underlying structure, and desires to validate or negate the hypothesized structure. The underlying structure is characterized by factor equations (Dillon & Goldstein, 1984, pp. 24, 53, 57-59; Long, 1983, pp. 11-15, 20).

Rogers and Kincaid examined social communication networks and used factor analysis to identify cliques. The observable indicators were the network links between individuals in one case and correlations from the communication matrix in a second case, and the factors were the cliques (Rogers & Kincaid, 1981, pp. 185-186, 188, 195).

2.3.1.2 MultiDimensional Scaling

Multidimensional Scaling (MDS) is a data reduction technique designed to take a data set, uncover its hidden structure, and represent the result pictorially. MDS is a mathematical tool for mapping objects in multidimensional space so the objects' relative positions in the space indicate their proximity, in a metric or nonmetric sense. Hence given a data set of distances (representing object similarities), MDS techniques reverse engineer the data to determine a graphical structure of the objects' relationships. Such a problem can become rather difficult when the data contains error, i.e. noise (Dillon & Goldstein, 1984, pp. 107-108, 125). Ji and Zha (2004) developed an algorithm, that incorporated MDS, for determining the physical positions of nodes in a wireless adhoc sensor network (Ji & Zha, 2004).

2.3.1.3 Cluster Analysis

Techniques designed to separate a data set into groups fall under the heading of cluster analysis. Additionally, the goal of cluster analysis is to find groups that exhibit small intra-group variation compared to the inter-group variation. Cluster analysis procedures examine data and discover groups based on datum similarity. Discriminant analysis, which is a dependence method, starts with the assumption that certain groups exist and attempts to classify data into groups (Dillon and Goldstein, 1984, pp. 157-158, 360). Consequently, discriminant analysis is looking for ways to distinguish between these groups, i.e. what makes them dissimilar. Cluster analysis techniques have been used to reconstruct relationships (i.e. trees) from species-related data (Vingron, Stoye, & Luz, 2002).

2.3.2 Dependence Methods

2.3.2.1 Multiple Regression Analysis

Multiple regression analysis deals with estimating and/or predicting a dependent variable's mean value on the basis of known (or fixed) values of (possibly) multiple explanatory/predictor variables. The (population) bivariate regression model postulates that $E(Y|X_i) = \beta_0 + \beta_1 X_i$, where Y is the dependent variable and X_i is an independent, predictor variable. However, it is reasonable to assume there exists some error between $Y = Y_i$ and $E(Y|X_i)$; consequently, the regression model $Y_i = E(Y|X_i) + \epsilon_i$. The error term ϵ_i represents (intentionally or unintentionally) excluded variables that nevertheless affect the dependent variable (Dillon & Goldstein, 1984, pp. 209-211). Consequently, hidden variables can be accounted for in the error term of the model; therefore, a large error term (variance) can indicate a hidden variable. The hidden variable (and proposed interactions with known variables) could be included to determine if the detailed model is more plausible (Myers & Montgomery, 2002, pp. 3-4; Robins, personal communication, April 2006).

The general multiple regression model with $p-1$ independent variables, X_2, X_3,..., X_p is expressed as $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_p X_{pi} + \epsilon_i$, $i = 1, \ldots, n$ where β_1 represents the intercept, and $\beta_2, \ldots, \beta_p$ denote the partial regression slope coefficients, and the residual term for the ith observation is labeled ϵ_i. Note that the independence assumption (i.e. orthogonality) nullifies any possible interaction between the explanatory variables (Dillon & Goldstein, 1984, pp. 209, 214-214).

2.3.2.2 Logistic Regression

Logistic regression is a model for analyzing the relationship of a dichotomous (or polytomous) dependent variable, D, with multiple independent variables, $X_i, i \in \mathbb{Z}^+$. If there exist n independent variables, then the logistic model for a dichotomous dependent variable is

$$P(D = 1|X_1, X_2, \ldots, X_n) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}}$$

where the α and β_i terms are parameters from the initial regression equation (Kleinbaum & Klein, 2002, pp. 5, 8, 268). Logistic regression has been used to identify network elements, e.g. Goldenberg, Kubica, and Komarek (2003), and an overview of such applications is provided in another section of this chapter.

2.4 Link Analysis

According to Harper and Harris (1975), "Link analysis methods were developed to systematically establish the relationships that exist among individuals and organizations from bits and pieces of available evidence" (p. 158). To some extent, link analysis can be viewed as inducing or learning network structures such as Bayesian networks (e.g. Cooper & Herskovits, 1992; Jensen, 2001, p. 36; Murphy, 1998). Nevertheless, there is a distinction between link analysis and other methods for constructing networks such as Bayesian networks. The latter methods find and characterize associations based on the overall statistical features of a sample of realizations gathered from some population. Link analysis, however, starts with network representable data and tries to gain insight from the network links and nodes (Association for the Advancement of Artificial Intelligence, 1998). The following paragraphs outline various efforts related, in some measure, to link analysis and this dissertation.

Kleinberg (1999) used link structure from the World Wide Web (WWW), instead of web page content, to analyze the set of pages germane to a broad search topic, and detect related 'authoritative' pages. He developed the concepts of authorities and hubs, and used linear algebra to identify the authoritative pages. Gibson, Kleinberg, and Raghavan (1998) employed Kleinberg's technique to infer web communities, resulting in a sense of ordered structure at a macro level of the WWW. Flake, Lawrence and Giles (2000) defined web communities graph-theoretically, and employed a maximum flow - minimum cut paradigm to identify the communities. Cai, Shao, He, Yan, and Han (2005) stated that entities are often related to each other in different ways depending upon the relation of interest, and implemented both regression and an algorithm based on the minimum cut concept to identify hidden communities in such a multiple network context (Cai *et al.*, 2005, pp. 58, 60-61). Gruhl, Guha, Liben-Nowell, and Thompkins (2004) developed an algorithm to infer a social network, containing directed arcs representing influence, from topic transmission in blogspace. Their propagation model is based on a stochastic cellular automata model employed by Goldenberg, Libai, and Muller (2001), who studied communications via word-of-mouth. Stochastic cellular automata models simulate higher level, i.e. system, effects from lower level entity interactions (e.g. Goldenberg *et al.*, 2001, pp. 213, 215; Gruhl *et al.*, 2004, pp. 491-493, 497-498; Liben-Nowell, 2005, pp. 90, 97-98; Ulam, 1952, p. 274; Von Neumann, 1966, pp. 91-92, 111, 133, 135). Kempe, Kleinberg and Tardos (2003) discussed the propagation model used by Goldenberg *et al.* (2001), when Kempe *et al.* (2003) examined the problem of maximizing influence propagation in social networks (note: this problem was presented in Domingos and Richardson (2001)); however, their effort was not focused on identifying the influence structure (or missing elements).

Cohn and Hofman (2001) developed a probabilistic model that incorporated both content and connectivity (i.e. link) details, and could be used to predict connectivity from content, e.g. link structure from content. Their domain was a collection of documents, but the concepts could be applied to social networks (Cohn and Hofman, 2001, p. 433). Kubica, Moore, Schneider, and Yang (2002) developed a method employing demographic information for determining group membership, and subsequent link data sampling with noise to identify group members that interact (Kubica *et al.*, 2002, pp. 798, 800). In order to more quickly identify groups, Kubica, Moore, and Schneider (2003a) developed a follow-on heuristic, similar to a k-means approach, to find groups based only on the link data. Experimental results were favorable, in that the heuristic found comparable solutions to the full method in a much shorter period of time. Additionally, Kubica, Moore, and Schneider (2003b) addressed the problem of learning a graph-based model from noisy, observed link data, that includes multiple link types and temporal information indicating the observation time. Additionally, the authors used their modeling approach, on real world data sets, to predict an entity's future links, even to entities with which there had been no former relations (Kubica *et al.*, 2003b, pp. 392, 399). Komarek (2004) experimented with logistic regression to determine if an object of interest is part of a given link (i.e. a relation of objects). Komarek's variants of logistic regression performed relatively well against other algorithms for the given task (Komarek, 2004, pp. 32, 94, 106). There have also been some accomplishments in detecting hidden groups (for a fixed number of actors) in social communication networks, with dynamic links, using Hidden Markov Models and random graph concepts (Baumes, Goldberg, Magdon-Ismail, and Wallace, 2004; Magdon-Ismail, Goldberg, Wallace, and Siebecker, 2003).

The above efforts primarily focus on discerning networks or groups; whereas, a significant portion of the research in this dissertation assumes defined groups, and attempts to determine unknown elements within the groups. The following paragraphs provide an overview of link prediction, and the relation to this dissertation.

According to Popescul and Ungar (2003), link prediction can include the following scenario: Given a set of nodes, some of the links between nodes, and possibly some node attributes; predict links that have not been observed. Consequently, they applied structural logistic regression for link prediction in the context of relational data (Popescul & Ungar, 2003, p. 92). Taskar, Wong, Abbeel, and Koller (2004) also performed link prediction in relational data scenarios (e.g. social networks), but they implemented the relational Markov network framework of Taskar, Abbeel, and Koller (2002). The relational Markov network extends a traditional Markov network by addressing relational data domains, and was employed by Taskar *et al.* (2002) to aid in entity classification, from a set of labels, in a relational data context (Taskar *et al.*, 2002, p. 485; Taskar *et al.*, 2004, p. 660). The modeling approaches of both Taskar *et al.* (2002) and Taskar *et al.* (2004) are based on the approach and concepts of Lafferty, McCallum, and Pereira (2001). Taskar *et al.* (2004) incorporated both attributes and link structure information (e.g. transitivity patterns), resulting in improvements in accuracy over flat classification, which attempts to predict links using only attribute information (Taskar *et al.*, 2004, pp. 659, 661). In a similar effort, Taskar, Abbeel, Wong, and Koller (2003) combined the approaches and concepts of Taskar *et al.* (2002) and Taskar *et al.* (2004) to address the problem of predicting labels (i.e. classifications) and links in relational data. Furthermore, Getoor, Friedman, Koller, and Taskar (2002) performed experiments that showed incorporating link structure in a model can provide better predictions of attributes (Getoor *et al.*, 2002, pp. 679, 700). For this dissertation, the training and testing procedures for identification and detection of group entities were, to some extent, derived from the concepts in Popescul and Ungar (2003) and Taskar *et al.* (2003).

Taskar *et al.* (2003) also discussed introducing hidden variables. While this dissertation contains concepts seen in the literature, the approaches for revealing nodes are somewhat different from, or not fully synthesized and implemented in, previous efforts. Additionally, the context and focus of this research are, to some extent, not identical to those encountered in the literature.

Goldenberg *et al.* (2003) examined an interesting facet of link analysis known as link completion, which they defined as determining the (most probable) missing entity or entities within a set of related entities, i.e. a link. The authors specifically examined various algorithms for the link completion case where only one member is missing. Collaborative filtering is defined as the use of a repository of other users' preferences, i.e votes, to predict items a new user might find appealing/useful (Breese, Heckerman, & Kadie, 1998, pp. 1-2; Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994, p. 175). Consequently, Goldenberg *et al.* (2003) noted the similarity of their examined case to the collaborative filtering protocol *Allbut1*, where a single vote of a user is randomly withheld, and then an attempt is made to predict the value of the vote based on the remaining votes of the user (Breese *et al.*, 1998, pp. 9-10; Goldenberg *et al.*, 2003, p. 2). The research presented in this dissertation contains differences from and a similarity to the problem of link completion. In one case of this dissertation research, link information for the missing node is not given. In fact, it is not assumed there exists a missing group node. The focus is on determining if there exists a missing node, not on selecting a candidate missing node from a set of nodes. Nevertheless, link completion is somewhat similar in objective to the reconstruction conjecture of Ulam (1960), discussed in another section of this chapter.

Hammer (1979/1980) examined the effects of distance, interaction level, and common connections between individuals on the formation, persistence and dissolution of links between individuals at two points in time. The general trend observed was that distance (measured by traversing individuals rather than physical lengths) is inversely related to the probability of a direct link, while intensity and common connections are directly proportional to the presence of a direct link (Hammer, 1979/1980). Liben-Nowell and Kleinberg (2004) examined the link prediction problem as it relates to social networks. The authors evaluate multiple measures, based only on network topological properties observed in a specified time interval, for their effectiveness in predicting links that will arise in a subsequent time interval. While certain measures outperform a random predictor, none of the measures achieve higher than a 16% accuracy (Liben-Nowell & Kleinberg, 2004, pp. 1, 18). Al Hasan, Chaoji, Salem, and Zaki (2006) combined features based on entity proximity, aggregate entity attributes and network topology to implement within several classification algorithms for link prediction in a co-authorship context. Additionally, the authors compared the performance of the features in addressing the task of link prediction, and shortest distance showed promise among the topological features (Al Hasan *et al.*, 2006, pp. 1, 3-5, 8-9). Adafre and Rijke (2005) attempted to identify missing links in Wikipedia web pages. The basis of their algorithm was that similar web pages (identified by content clustering) should have similar link structure, i.e. in effect structural equivalence (Adafre & Rijke, 2005, pp. 90, 92, 94). While both this dissertation and the efforts reviewed in this paragraph address network elements, the former can address nodes, while the latter focuses on links.

Cooke (2006) makes a distinction between link prediction and link detection; the former is concerned with identifying new links in successive time steps (per Liben-Nowell and Kleinberg (2004)), while the latter deals with detecting missing links in the current time step. To simulate a hidden link in the detection task, Cooke (2006) removed links of connected dyads. He then examined the effectiveness of various metrics in detecting/distinguishing hidden links versus links that were either forming (i.e. predicted) or links that were truly non-existent. Hidden and forming links could not be distinguished (at the author's prescribed level of confidence) using traditional topological metrics; however, for simple networks, different link types were identifiable when various metric types (e.g. temporal) developed by Cooke (2006) were employed (Cooke, 2006, pp. 60, 63, 68, 108, 110-111). Hoff and Ward (2006) implemented a latent factor model to analyze link detection and prediction, and multi-way analysis of link data. The detection and prediction results were promising compared to a random modeling approach (Hoff & Ward, 2006, pp. 3-4, 7-12). While there is similarity in the node detection approach of this dissertation to the link detection technique in Cooke (2006), the fundamental difference is the type of network element, i.e. a node vice a link.

2.5 Causality

The previous data analysis methods involve correlations and covariations among variables. However, such relationships do not necessarily imply causality. Yet causality is important to understand since certain networks and systems, especially those that are rigorously designed and engineered, are by nature causal (National Research Council, 2005, pp. 7, 11-12). This section reviews methods that attempt to identify causal relationships among variables; furthermore, examples of literature pertinent to ascertaining hidden causes are provided.

2.5.1 Structural Models

Structural models are representations of theories or theoretical models, where the term theories refers to sets of causal hypotheses that attempt to explicate phenomena occurrence (Singh, 1975; James, Mulaik, & Brett, 1982, pp. 27, 31, 68, 104-105). A structural model is represented as a simple, directed graph (cf. Figure 1). In such a graph, the nodes represent causes, effects or error terms. Causes that are hypothesized and hence not to be explained are referred to as exogenous variables, while effects (to be explained by the model) are labeled endogenous variables. Note that it is possible for an effect to, in turn, cause another effect; however, in such a case, both the original effect and the succeeding effect are considered endogenous variables. The error terms can represent 'disturbances' on the endogenous variables such as omitted causes, and random/non-random measurement errors. Straight arcs representing a causal relationship may originate from any of the three node types, but terminate only at endogenous variables. Furthermore, each straight arrow has a structural parameter associated with it, indicating the strength of the causal relationship. Double-headed curved arcs occur between the exogenous variables, i.e. causes, indicating the existence of a relationship (not to be explained by the model) between such variables. Figure 1 illustrates a structural model with exogenous variables X_1 and X_2, and endogenous variables, Y_1 and Y_2. The B terms represent structural parameters and the d terms denote disturbances. Another aspect of variables is whether they are associated with directly observable and consequently measurable events, in which case they are referred to as manifest variables, or associated with events not directly observable and so not directly measurable, i.e. latent variables. If a variable is latent, it is indirectly measured via associated manifest variables (cf. factor analysis) (James *et al.*, 1982, pp. 31-33, 55).

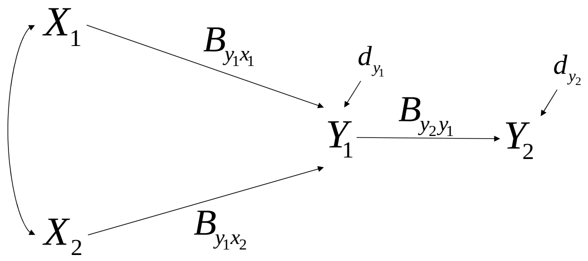


Figure 1 Illustrative Structural Model; adapted from James *et al.* (1982)

Mathematically, a structural model can be represented by a set of functional equations, with the effect as the dependent variable, and causes and an error term as the independent variables. Often, the functional relation is deemed a linear combination of the causes and the error terms, and the causes have parameters indicating causal strength. Such a model is referred to as a linear causal model. Consider the situation where a variable, i.e. cause, is unmeasured (because it is not deemed as part of the hypothetical causal structure) but is actually relevant to the causal structure. The result of such a situation is that the effect's error term will include the unmeasured cause, and be related to at least one of the measured causes, i.e. exogenous variables. Consequently, there will exist covariation between the effect's error term and relevant measured cause(s) of the effect; however, James *et al.* (1982) argued that estimating or solving for such covariation is not possible since disturbance terms, by definition, cannot be directly measured (James *et al.*, 1982, pp. 22-23, 33, 44-47, 62-65).

A path model is a structural model with manifest variables in standardized form, and the structural parameters are denoted as path coefficients. Additionally, the associated functional equations and confirmatory analysis are referred to as path equations and path analysis, respectively (James *et al.*, 1982, pp. 68-69).

Latent variable (structural) models are generalized forms of linear causal models in which some of the variables are unobserved. A measurement model is a model in which manifest variables (effects) serve as indicators of the latent variables (causes), and the causal relations between the two variable types are specified. Note that measurement models may include disturbance terms (James *et al.*, 1982, pp. 104, 106-107). Silva, Scheines, Glymour and Spirtes (2006) defined a pure measurement model as, "a measurement model in which each observed variable has only one latent parent, and no observed parent" (p. 198). With respect to such models, Silva *et al.* (2006) developed methods for, "discovering which latent variables exist, and which observed variables measure them...[and]...discovering the Markov equivalence class that contains the causal graph connecting the latent variables" (p. 193).

2.5.2 Latent Class Analysis and Models

A latent class model is a subset of latent structure models, which are related to factor analysis and structural equation models; however, a latent class model assumes the manifest variables are categorical (vice continuous) and indicate latent variables that are also categorical, i.e. composed of various classes. The objective of latent class analysis is to characterize the latent variable that explicates the association observed between manifest variables (Dillon & Goldstein, 1984, pp. 490-493; Lazarsfeld & Henry, 1968, pp. 11, 17, 21, 46).

Hierarchical latent class models are Bayesian networks whose structures are rooted trees with observable leaves, but the remaining nodes are latent. Zhang (2004) developed an algorithm for learning hierarchical latent class models, building the structure up from the original latent class model, in part, by introducing a new hidden variable as a parent of two observable variables that violate the assumption of local independence (i.e. observable variables are mutually independent given the original hidden variable) (Goodman, 1974, p. 1179; Zhang, 2004).

2.5.3 Bayesian Networks and Causal Inference

Markov networks are undirected graphs where symmetrical probabilistic dependencies are represented by links. Bayesian networks are DAGs with nodes and links representing random variables and direct causal influences (measured via conditional probabilities), respectively (Pearl, 1988, pp. 50-51, 77, 90-91, 96). Furthermore, a dependency model is a rule that determines if the assertion, "Given **C**, **A** and **B** are independent", i.e. $I(\mathbf{A},\mathbf{C},\mathbf{B})_M$, is true. A probability distribution is an example of a dependency model, since one can determine the independence of two variables given a third variable via conditional independence upon the third variable. In attempting to graphically represent (via an undirected graph) a dependency model, a few more definitions are necessary. A dependency map (D-map) is an undirected graph representing the dependency model such that any independence in the model, i.e. $I(\mathbf{A},\mathbf{C},\mathbf{B})_M$, implies that the subset of graph vertices corresponding to **C**, intercept all paths between the graph vertices in the subsets **A** and **B**. An independency map (I-map) embodies the converse implication, and a perfect map incorporates both implications. Given these definitions, it is possible to define a Markov network of the dependency models as a minimal I-map, i.e. deletion of any associated graph edge would make the graph no longer an I-map (Pearl, 1988, pp. 83, 91-92, 96). In determining the dependency graph from empirical data, one can also incorporate outside information to reduce the computations that must be performed (Shipley, 2002, pp. 253-254, 258-259). Such outside information for a social (i.e. influence) or engineered network could be organizational construct constraints or the laws of physics, respectively (National Research Council, 2005, pp. 7, 12).

Cooper and Herskovits (1992) derived a Bayesian method for inducing probabilistic networks, specifically Bayesian belief networks, from data. Their method addressed the comparison of probabilities for different (sub)network topologies (both directed and undirected). Consequently, one can derive the probability of an arc's existence; furthermore, in the context of causal models, information regarding the likelihood of a causal relationship can be derived (Cooper & Herskovits, 1992, pp. 309, 312, 318-320). This latter feature is beneficial when attempting to address causality using small sample sizes (Cooper & Herskovits, 1992, p. 318; Pearl, 2001, p. 64). A limitation of the author's method is its set of assumptions, e.g. *a priori* knowledge of the (temporal) ordering of variables (Cooper & Herskovits, 1992, p. 320; Korb & Nicholson, 2004, p. 200). The effort of Cooper and Herskovits (1992) is a natural predecessor to the work of Butts (2003) involving a Bayesian approach to recreate the criterion graph based on informant reports (Butts, 2003, pp. 105-106). Cooper and Herskovits' research also addresses determining the most probable network structure. They provide an exact method (using some assumptions) and a heuristic. Additionally, they identified open problems related to efficiently searching and calculating probabilities, to include situations with hidden variables (Cooper & Herskovits, 1992, pp. 318, 320-321, 323, 326, 335-336).

Dynamic Bayesian Networks (DBNs) are formed by taking a series of Bayesian networks over time. Typically, the Bayesian network structure, at each time slice, remains the same. Furthermore, the arcs connecting Bayesian networks at different time slices (i.e. inter time-slice arcs from a node in one Bayesian net to another) are only between networks in adjacent time slices, thus the Markovian property is preserved. Evidence introduced about a set of nodes in a particular time slice, can be used to update distributions in other nodes, to include those in future time slices. (Boyen, Friedman, & Koller, 1999, pp. 97-98; Dean & Kanazawa, 1989, p. 148; Friedman, Murphy, & Russell, 1998; Kjærulff, 1995, pp. 91-92; Korb & Nicholson, 2004, pp. 105-106; Murphy & Mian, 1999). Dean and Kanazawa (1989) argue that such models permit thinking about planning applications and computing associated probabilities in a manner more direct than equivalent Markov models (Dean & Kanazawa, 1989, p. 148). As indicated by Murphy and Mian (1999), various authors have addressed hidden variables for DBNs in a number of ways.

The following efforts highlight the methods used to detect the presence of hidden, i.e. latent, variables. Pearl (1986) showed that if a tree-structured representation of a Bayesian network exists, then it is possible to uniquely uncover the topology of the tree (to include its hidden variables, i.e. internal tree nodes) by observing pairwise dependencies among the observable variables, i.e. the tree leaves (Pearl, 1986, pp. 241, 273-274, 277). Similarly, Zhang, Nielsen and Jensen (2004) developed an algorithm to learn network tree structures that incorporated hidden variables indicated by the violation of mutual independence between feature, i.e. observable, variables given a class variable (Zhang *et al.*, 2004, pp. 283-284, 288-289).

Cooper and Herskovits (1992) developed methods to address missing variable values and hidden variables (i.e. no data available); however, their methods are not efficient for practical usage. This occurs since the authors' formulation for computing the probability of a network structure has an exponential complexity with respect to the missing values. Another result provided by the authors is the expectation of a conditional probability over all possible network structures for a set of variables, permitting inference by averaging multiple belief networks' inferences. A possible extrapolation of this technique is the derivation of a system of systems inference from multiple subsystem network structure inferences (Cooper & Herskovits, 1992, pp. 309, 323-328).

Connolly (1993) also developed a method for detecting and inserting hidden nodes in the construction of Bayesian network trees. His method involved measuring the dependence of observable variables (e.g. X and Y) using the mutual information formula:

$$\sum_i \sum_j P(X_i, Y_j) \log \frac{P(X_i, Y_j)}{P(X_i)P(Y_j)}$$

and then clustering highly dependent variables (Connolly, 1993, p. 66).

Martin and VanLehn (1994) developed Bayesian network topologies, with hidden variables, having a factor structure. The authors detected hidden variables (i.e. factors) and introduced such variables into the topology by finding cliques of dependent (based on Pearson's χ^2 association test) observable variables (Martin & VanLehn, 1994, pp. 2-4). Elidan, Lotner, Friedman and Koller (2000) looked for structural signatures, i.e. semi-cliques, in order to identify and introduce hidden variables (Elidan, Lotner, Friedman, & Koller, 2000, pp. 479-480). The semi-cliques addressed by Elidan *et al.* (2000) are similar to communication network cliques discussed in Rogers and Kincaid (1981). Elidan *et al.* (2000) introduced hidden variables within the semi-clique; whereas for communication networks the hidden variables could be either intra-clique in the case of bridges, or between cliques in the case of liaisons (Elidan *et al.*, 2000, p. 482; Rogers & Kincaid, 1981, pp. 346-347). Friedman (1997) developed a method, based on the Expectation-Maximization algorithm of Dempster, Laird and Rubin (1977), to learn the structure of Bayesian networks, given incomplete data. Friedman (1997) defined incomplete data as either hidden variables or missing variable values (Friedman, 1997, p. 125).

A significant milestone in the analysis of causal models was reached when Geiger and Pearl (1990) provided a theorem that allowed for a translation between DAGs and probability distributions. Before elaborating on the consequence of their theorem, it is necessary to discuss the notion of *d*-separation. *d*-separation is a criterion for determining if a variable or a set of variables in a causal model are independent of another variable or set of variables. When the model contains only undirected links (hence no longer a causal model, i.e. a Markov network), then *d*-separation is equivalent to identifying vertex cut-sets (Geiger & Pearl, 1990, pp. 3, 10; Pearl, 1988, pp. 88, 93-94, 116-117; Pearl & Paz, 1987; Shipley, 2002, pp. 23, 29). Pearl (1988) formally defines d-separation as follows:

> If **X**, **Y** and **Z** are three disjoint subsets of nodes in a DAG D, then **Z** is said to d-separate **X** from **Y**, denoted $<\mathbf{X}|\mathbf{Z}|\mathbf{Y}>_D$, if there is no path between a node in **X** and a node in **Y** along which the following two conditions hold: (1) every node with converging arrows is in **Z** or has a descendant in **Z** and (2) every other node is outside **Z**. (Pearl, 1988, p. 117)

According to Pearl (1988), a DAG D is an I-map of a dependency model M if $<\mathbf{A}|\mathbf{C}|\mathbf{B}>_D \implies I(\mathbf{A},\mathbf{C},\mathbf{B})_M$, where the subscripts D and M refer to the DAG and dependency model, respectively. Furthermore, D is called a Bayesian network if and only if D is a minimal I-map of a probability distribution P (that is on a set of variables) (Pearl, 1988, p. 119). Therefore, the essence of Geiger and Pearl's theorem is that for any DAG, there exists a probability distribution that incorporates all the independencies shown in the causal model. Consequently, if one observes statistical independencies in the observation data that are not portrayed in the hypothesized causal model (i.e. DAG), then the model is incorrect. Note: The independencies in the causal model are identified via the notion of *d*-separation (Geiger & Pearl, 1990, pp. 3, 10; Pearl 1988, pp. 116-117, 119, 122; Shipley 2002, pp. 36-37; Verma & Pearl, 1990, p. 71).

Pearl and Verma (1991) developed an algorithm for inferring causal models from observations (even with the presence of some unobservable variables and without temporal information). Their algorithm searches for conditional independencies between a pair of variables, assumes a causal theory probability distribution is available, and attempts to derive the causal model's topology from the distribution's features; however, Pearl and Verma (1991) argued that a large sample is a sufficient proxy for the true distribution. Their algorithm was designed to distinguish between genuine causes and spurious covariations, i.e. variables with a hidden common cause. Additionally, Pearl and Verma (1991) noted that their modeling construction task was an identification game. This is similar in thought to the identification problem of reconstructability analysis (discussed in another section) mentioned by Cavallo (1980), and Cavallo and Klir (1981) (Cavallo, 1980, pp. 647-648; Cavallo & Klir, 1981, p. 2). The original algorithm was given in Verma and Pearl (1991), who discussed how (topologically) different causal models can be statistically equivalent; consequently, an equivalence class of models can result which the authors succinctly represent by a graphical representation known as a pattern (either rudimentary or completed). Verma and Pearl (1991) also addressed the notion of an embedded causal model (i.e. a model in which not every variable is observable) and associated patterns, represented by graphs containing both singly directed and bi-directional links. Subsequently, they provide a boundary, given a number of observable variables, on the number of distinct embedded causal models (Verma & Pearl, 1991, pp. 255-256, 259-261, 263).

Glymour *et al.* (1987) provided a heuristic for introducing latent variables into a causal model. The heuristic is predicated on searching a (sample) correlation matrix for satisfaction of tetrad constraints (i.e. differences of correlation products equivalent to zero). Partial correlation constraints are also implemented in the authors' approach and aid in determining placement of the latent variable(s) (Glymour *et al.*, 1987, pp. 75-86, 178-183). In a follow-on article, Glymour and Spirtes (1988) illustrated that it was possible to derive more than one causal structure, satisfying the constraints, from the empirical data; furthermore, their method was limited to models with linear causal relations. Glymour and Spirtes (1988) also addressed the issue of model specification for time-series models, but the question of latent variables for such models was open at the time the article was written. Spirtes, Glymour, and Scheines (1990) tied together their previous results and Pearl's (1988) d-separation principle; consequently, a sufficient condition for the presence of latent variables in causal models was provided (Spirtes, Glymour, & Scheines, 1990).

Spirtes, Glymour, and Scheines (1993, 2000) provided an algorithm for inferring a causal graph even in the presence of latent common causes. The algorithm, causal inference (CI), took a covariance matrix or cell counts as input, and output a special type of partially oriented graph known as a partially oriented inducing path graph. Hence, the exact causal structure may not be known (e.g. arrow directions), but at least the number of possible structures is (somewhat) reduced (Spirtes, Glymour, & Scheines, 1993, pp. 180-183; Spirtes, Glymour, & Scheines, 2000). Unfortunately, the CI algorithm is not computationally feasible when there are many variables; therefore, Spirtes *et al.* (1993, 2000) developed a fast causal inference (FCI) algorithm that also outputs a partially oriented graph (Glymour, Scheines, Spirtes, & Ramsey, 2004a; Richardson, 1996; Spirtes *et al.*, 2000). This algorithm is feasible for large numbers of variables if the true causal structure is sparse and there are only a few bi-directional edges chained together. Furthermore, Spirtes *et al.* (1993, 2000) noted that their FCI algorithm can, in certain cases, provide more orientation information than Verma and Pearl's IC algorithm. Note: both the CI and FCI algorithms implement d-separation (Spirtes *et al.*, 1993, pp. 171, 182, 188-189, 200;

Spirtes *et al.*, 2000, pp. 129-130, 139-140, 144-145, 155). The FCI algorithm assumes the Causal Markov and Faithfulness conditions hold (Spirtes *et al.*, 2000, p. 124). The causal Markov condition is satisfied by a causal graph with vertex set $\mathbf{V}$ and a probability distribution (generated by the causal structure represented by the graph) if and only if every vertex in the graph is independent of $\mathbf{V} \setminus (\mathbf{Descendants}(W) \cup \mathbf{Parents}(W))$ given $\mathbf{Parents}(W)$, where W is any vertex in the graph (Spirtes *et al.*, 2000, p. 29). The Faithfulness condition is satisfied by a causal graph and a probability distribution (generated by the graph) if and only if every relation of conditional independence that is true in the probability distribution is implied by applying the Causal Markov condition to the graph (Spirtes *et al.*, 2000, p. 31).

Lemmer (1996) presents an alternative perspective to account for correlation between variables. He argues that causes generate signals to which effect-events observe and respond. Thus a latent cause should be introduced between the original cause and the effects, such that the states of the latent variable are the cross product of the signal states issued by the original cause. Lemmer (1996) shows that such a representation is more efficient, with respect to storage, than a common approach to latent variable introduction, where the variable is added at the same level as the original cause (Lemmer, 1996, pp. 7, 13-14).

Regardless of the implementation, causal exploratory analysis is concerned with generating candidate causal structures (Dillon & Goldstein, 1984; Shipley, 2000; Spirtes *et al.*, 2000). The next section reviews literature concerned with a slightly different problem; specifically, determining the true structure of a network's graph representation from partial information of the structure. This reconstruction process can be applied to causal or non-causal networks.

2.6 Network Reconstruction

Reconstructing a structure from its components is important in a variety of fields, e.g. link analysis in police intelligence (Harper & Harris, 1975). The following sections provide an overview of various methods that can address network reconstruction.

2.6.1 Reconstructability Analysis

Cavallo and Klir (1979) use the term reconstructability analysis (RA) to refer to the process of examining the possibilities of reconstructing desirable properties of overall systems using knowledge of respective properties of their various subsystems (Cavallo & Klir, 1979, p. 143). Systems in RA can have causal (thus directed) relations (Klir, 1985, pp. 151-153; Zwick, 2004, p. 889). In their discussion of discovering causal structures, Spirtes *et al.* (2000) noted the reconstruction effort of Klir and Parviz (1986). Additionally, latent variable modeling can be implemented within the reconstruction problem context; however, Zwick (2004, 2007) notes that this capability does not appear to have been developed or applied in RA (Zwick, 2004, pp. 878, 883, 889; Zwick, personal communication, May 2007).

RA is composed of two problems, reconstruction and identification. The reconstruction problem attempts to determine the set of subsystems that can adequately reconstruct (in terms of behavior or properties) the known overall system. The identification problem addresses ascertaining what information can be gained about an overall (unknown) system, or set of systems known as the reconstructability family, from its subsystems (Cavallo, 1980, p. 648; Cavallo & Klir, 1981, p. 2; Klir, 1985, pp. 151-153, 212, 227-228). The identification problem is also concerned with choosing a single system from the reconstruction family as the hypothesized overall system. This subproblem can be addressed by establishing some goodness criteria and then performing optimization to determine the best system candidate(s). Since the true system is unknown, it is not possible to measure the difference between members of the reconstruction family and the true system; however, one can identify an unbiased reconstruction as an estimate of the true system. The unbiased reconstruction is the solution (i.e. system) that is based on all of the information in the subsystems, but no more than that information. According to Klir (1985), the unbiased reconstruction is a unique solution (for probabilistic systems) to the optimization problem known in the literature as the principle of maximum entropy (Cavallo, 1980, p. 648; Klir, 1985, pp. 222, 228). Consequently, the unbiased reconstruction system can be considered an initial solution that can possibly be augmented as related, substantive knowledge becomes available. Alternatively, one may choose a reconstruction to minimize risk (Klir, 1985, pp. 222-223).

2.6.2 Graph Reconstruction

In graph theory, a vertex-deleted subgraph of an undirected graph G is a subgraph with a single vertex (and its adjacent edges) deleted from G (Bondy, 1991, p. 221). There are $n(G)$ vertex-deleted subgraphs of G, where $n(G)$ is the number of vertices in G (West, 2001, p. 38). The (unlabeled) vertex-deleted subgraphs of G are referred to as cards, and the deck is considered the entire family of such subgraphs. A graph H, containing the same deck as G, is called a reconstruction of G. Furthermore, if every reconstruction of G is isomorphic to G, then G is reconstructible. The Reconstruction Conjecture states that all finite, simple (unlabeled) graphs with at least three vertices are reconstructible, i.e. G is unique up to an isomorphism (Bondy, 1991, p. 221-223; Harary & Manvel, 1970; Kelly, 1957, p. 968). Kelly (1957) proved an equivalent theorem for trees, and then verified the conjecture for simple graphs up to order seven (Kelly, 1957; Myrvold, 1990, pp. 150). Additionally, Ulam (1960) conjectured the problem using set notation (Bondy & Heminger, 1977, p. 249; Ulam, 1960). There are similar definitions for digraphs; however, in general such graphs are not reconstructible (Bondy, 1991, pp. 221, 223; Stockmeyer, 1977, 1981). Using edge-analogous concepts, Harary (1964) developed what is referred to as the Edge Reconstruction Conjecture; specifically, every finite simple graph with at least four edges is edge reconstructible. Additionally, Harary (1964) proposed the problem of reconstructing a graph from its non-isomorphic subgraphs. This is referred to as the set reconstruction conjecture (Bondy, 1991, p. 221; Harary, 1964, pp. 51-52; Lauri, 2004, p. 86).

The reconstruction number of a graph is the minimum number of vertex-deleted subgraphs of G necessary to identify the unique graph G (Harary & Plantholt, 1985; Myrvold, 1990). Harary (1964) posed the problem of determining the minimum number of vertex-deleted subgraphs necessary to reconstruct a graph, and Harary and Manvel (1970) showed that a graph with at most two unlabeled vertices (where the labels are distinct) is reconstructible from three of its vertex-deleted subgraphs (Harary, 1964, p. 51; Harary & Manvel, 1970, pp. 136, 143). Additionally, Harary and Manvel (1970) analyzed bounds for various graphs and noted that not every graph requires all of its vertex-deleted subgraphs to reconstruct the original graph (Harary & Manvel, 1970, pp. 133-136). Harary and Plantholt (1985) conjectured that almost all (unlabeled) graphs could be reconstructed with three vertex-deleted subgraphs (Harary & Plantholt, 1985, p. 454). Bollobás (1990) proved the conjecture from a probabilistic perspective using random graphs (Bollobás, 1990; Bondy, 1991, p. 222). Consequently, if one has multiple vertex-deleted subgraphs, perhaps representing longitudinal observations of a graph with a single vertex and its incident edges removed (e.g. a network with a single, hidden node), it may be possible to reconstruct the unique, original graph.

Myrvold (1988, 1990) has researched reconstruction with respect to ally and adversary numbers, where the terms ally reconstruction number and reconstruction number are synonymous. The concept is based on an ally providing subgraphs to an individual in an ordering to minimize the number required to reconstruct the original graph. Alternatively, the adversary reconstruction number is the number of subgraphs necessary in order to determine the original graph when an adversary is providing the ordering of subgraphs received (Myrvold (1992) noted that the adversary reconstruction number concept was mentioned in Harary and Manvel (1970)). Myrvold (1990) proved that the reconstruction number of a tree with five or more vertices is three (Myrvold, 1988; Myrvold, 1990, p 150).

While reconstruction has received significant attention, the literature does not appear to address the topic with respect to either repeat observations of the subgraphs nor causal applications/constraints. Incorporating these concepts into reconstruction is a contribution which this dissertation addresses.

2.7 Conclusion

This chapter has provided an overview of work related to this research. The primary areas of relevant literature include both analytic (to include graph-theoretic) methods and network models that can address incomplete data, as well as causal information. While the literature is replete with such information, there is room for synthesis and expansion of current problem formulations and solutions with respect to characterizing and detecting unrevealed elements in network systems. The next chapter contains the methodology that will be employed to address some of the related gaps.

3. Methodology

3.1 Introduction

The crux of this dissertation is the development and demonstration of methods to characterize and detect unrevealed elements in network systems. The associated manifestations are three-fold:

1. A method to identify (and consequently characterize) and detect individuals that bridge groups in social networks. In network terms, these individuals are referred to as connection nodes. The method is demonstrated on empirical and generated data.
2. A method to reconstruct network structures given repeat observations of various parts of the network structure. The method is exhibited on a few small sized graphs. Additionally, an approach to address the reconstruction of causal/influence networks is provided.
3. A method to determine possible social influence network (SIN) structures and hidden individuals using causality analysis. An application of the method using empirical data is demonstrated.

The networks are assumed to be representable by simple graphs (i.e. no loops), which may be directed or undirected depending on the context and the problem being addressed.

The following sections provide relevant concepts from the literature, approaches and assumptions for the contributions provided in Chapter 1.

3.2 Revealing Bridges in Social Networks

In this research, a bridge (member) is an element that connects two or more groups, and is an element of only one of the groups (Rogers & Kincaid, 1981, p.

29). If there is only one bridge member per group, the bridges can be graphically represented as articulation points, similar to the graphical representation of liaisons (Harary & Norman, 1953, p. 27; Ross & Harary, 1955, pp. 253, Weiss & Jacobson, 1955, p. 664). Liaisons are network elements linked to individuals in at least two groups other than their own group; furthermore, the liaison's group might consist only of the liaison (Weiss & Jacobson, 1955, pp. 664, 666). Hence a bridge is a type of liaison, and for this research it is assumed bridge members and liaisons are similar in attributes since both connect groups.

For this dissertation, revealing bridges is composed of two activities: identification and detection. Identification is classifying/labeling group members as either a bridge or non-bridge. Detection is inferring the existence of an unknown bridge from group data that does not contain the bridge in its membership, the bridge's attributes, or the bridge's links.

Given the potential distinguishing characteristics of liaisons mentioned in the literature review, the assumption that inter-group communication is necessary in order for a system to effectively function, and a situation where the groups are composed of highly homophilous elements, then it is reasonable to expect the existence of bridges that are somewhat heterophilous with respect to the known group entities (Ferrand *et al.*, 1999, pp. 204-205; Granovetter, 1973; Lenski, 1954, pp. 405-406; Liu & Duff, 1972; Rogers & Bhowmik, 1971, pp. 532-533; Rogers & Kincaid, 1981, pp. 128-129). Based on this idea of distinction, as well as, concepts and/or logistic regression and discriminant analysis assumptions and features given in Clark (2005), Dillon and Goldstein (1984), Hosmer and Lemeshow (2000), Kleinbaum and Klein (2002), Lachenbruch, Sneeringer and Revo (1973), Montgomery, Peck and Vining (2001), Neter, Kutner, Nachtsheim and Wasserman (1996), Pohar, Blas and Turk (2004), logistic regression was the method chosen in this research for identification. The dependent variable is dichotomous, and represents whether an individual is or is not a bridge. The covariates involve both human factor and structural components, consistent with theories and findings in the relevant literature. Consequently, identification permits characterization of the member according to the covariates in the regression model. It should be noted that the structural attributes are intra-group; otherwise, the identification of a bridge might be evident by inspection, e.g. a member with an inter-group link is a bridge. The identification method is validated using empirical and generated group data, partitioned into estimation and prediction components, with all member labels, e.g. bridge/non-bridge, in the prediction data removed. A logistic regression model, fit to the estimation data, is applied to the prediction data, and the classification results examined (Montgomery *et al.*, 2001). The data sets, and associated attributes and groups are described in the chapter demonstrating the presented methods.

The detection method is a logical extension of the identification method and the assumption that the group under analysis communicates with at least one other group in the network. Specifically, if the regression model does not identify a bridge in a group, then the existence of an unknown bridge member or members may be implied given the underlying assumption. Alternatively, it is possible the model may misclassify members, and such cases are examined, to some extent, in this research. Validation of the detection method is accomplished in a manner similar to the identification method. There is one additional step; specifically, the appropriate bridge (and its incident links) in the prediction data is removed prior to applying the logistic regression model that was fit to the estimation data.

Once a bridge node is detected, there exists the issue of placement within the group, i.e. to whom it is connected. A heuristic approach is to assign the average (or mode) of the non-bridges demographic values to the detected bridge, and then connect the detected bridge to any node with which the similarity of demographic attributes is 0.5 or greater. An optimization technique is to maximize the probability that the inferred individual equals a bridge subject to parametric/structural constraints. Both approaches are demonstrated and results provided in a subsequent chapter.

This section has provided an overview of the methods to reveal bridges in social networks. Yet as previously noted, there exist non-social network bridges that have analogous characteristics to social network bridges. Consequently, the above methods should apply to a non-social network.

3.3 Network Reconstruction

Network reconstruction refers to the following problem: Given a n node network (with unknown structure) represented by a graph, G_u, how many observations are required to reconstruct, if possible, G_u, and what is the accuracy of the reconstruction as observations are made under the following constraint: there are only enough resources to observe $n-1$ nodes with their associated links in a single observation (Kelly, 1957, p. 968). This problem definition was derived from the field of graph reconstruction. The Reconstruction Conjecture asserts that all finite, simple (unlabeled) graphs with at least three vertices are reconstructible, i.e. G is unique up to isomorphism (Bondy, 1991, pp. 221-223; Harary & Manvel, 1970; Kelly, 1957, p. 968). Traditional graph reconstruction operates under the assumption that the vertex-deleted subgraphs are provided, i.e. observed, in some order without repetition; while there may exist isomorphic vertex-deleted subgraphs, each of the n vertex-deleted subgraphs are observed once and only once. Furthermore, the focus of reconstruction has been the reconstructability of graphs; rather than reconstruction accuracy. Nevertheless, as demonstrated in this dissertation, concepts of ally and adversary reconstruction numbers can be employed to address accuracy and temporal aspects in an alternative reconstruction framework (Bondy, 1991; Harary, 1964; Harary & Plantholt, 1985; Lauri, 1987, 1992; Myrvold, 1988, 1990).

Bondy (1991) (based on personal communication with Stockmeyer in 1976) noted that observations of vertex-deleted subgraphs of a reconstructible graph, G_u, can imply more than one G_u, if all vertex-deleted subgraphs are not observed (Bondy, 1991, p. 222). For example, in the context of traditional reconstruction, consider G_u, G_{t_0} and G_{t_1} as shown in Figure 2. G_{t_0} and G_{t_1} represent two different vertex-deleted subgraphs of G_u observed at time t_0 and t_1, respectively. While G_{t_0} and G_{t_1} appear identical in structure, they are 'different' because two different vertices were deleted from G_u to produce G_{t_0} and G_{t_1}. From these vertex-deleted subgraphs, two representations of G_u are G_1 (which is equivalent to G_u) and G_2, as displayed in Figure 3. Results for this problem area are provided in another chapter, but it is clear that ambiguity in traditional network reconstruction can arise.

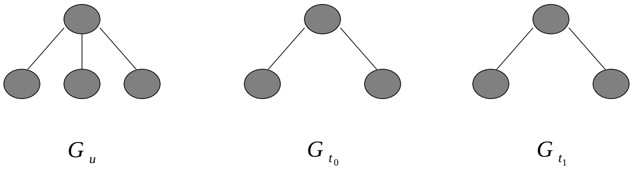


Figure 2 Undirected Graph with First Two Time Steps

This dissertation extends network (i.e. graph) reconstruction by developing and demonstrating a framework in which reconstruction is attempted and analyzed with repeat observations of the vertex-deleted subgraphs permitted; conceptually derived, in part, from random graph (and associated evolutionary) concepts of Bollobás (2001), and Erdős and Rényi (1960), as well as, (ally and adversary) reconstruction

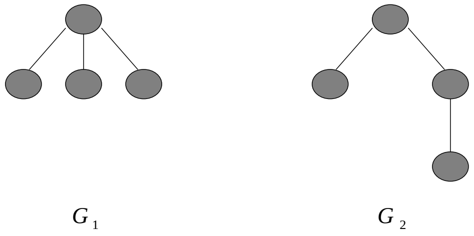


Figure 3 Possible Reconstructed Networks

number concepts of Harary (1964), Hararay and Plantholt (1985), Myrvold (1988, 1990) (Bollobás, 2001, p. 42; Erdős & Rényi, 1960, p. 20). Consequently, the level of ambiguity/difficulty associated with reconstruction (with or without observing all vertex-deleted subgraphs) can increase since repeat observations of the vertex-deleted subgraphs are permitted. The extension is approached as follows:

1. Framework definitions and notation are developed.
2. A general formula for reconstruction accuracy is given.
3. The framework is demonstrated on a simple, undirected graph with unlabeled vertices, and related assertions are provided.

Another paradigm considered in this dissertation is the incorporation of causal exploratory analysis within network reconstruction. In this case, the underlying network represents a causal/influence network, which is represented as a DAG. The approach for this situation is similar to the previous reconstruction approach.

3.4 Determining SIN Structures and Hidden Individuals

In this dissertation, a social influence network (SIN) is defined as a network of individuals who may exert influence on one another. Based on concepts in the literature, it is assumed a SIN can be represented by a causal network; consequently, a SIN will be portrayed as a directed acyclic graph (DAG). In certain situations, it may be appropriate to assume the SIN nodes act in concert for some purpose; therefore, the SIN could be portrayed as a connected DAG. While the focus in this section is on social influence networks, the methods presented can be applied to a variety of other networks that operate on the principle of a node causing effects on or influencing another node or nodes (National Research Council, 2005, pp. 7, 11-12; Pearl, 1988, 2000).

Given the assumption that a SIN can be represented by a causal network, the method implemented to determine possible SIN structures is causal exploratory analysis. Such analysis requires identifying a measurement of causality/influence so that probabilistic independencies between the nodes can be calculated (Geiger & Pearl, 1990, pp. 3, 10; Pearl, 1988, pp. 81-86, 89 91-94, 116-119, 122; Shipley, 2002, pp. 8, 9, 36-37, 90-94; Spirtes *et al.*, 2000, pp. 43-44, 82, 139; Verma & Pearl, 1990, pp. 71; Verma & Pearl, 1991, pp. 256, 264). Independency calculations between nodes of causal models representing some network types (e.g. engineered networks) are, generally, not too difficult because such networks have nodes that can naturally be represented in an event/state form that is measurable. Pearl (2001) gives an example of a Bayesian network containing a node representing the states (on and off) of a sprinkler (National Research Council, 2005, pp. 11-12; Pearl, 1988, pp. 18-19, 50-51; Pearl, 2001, pp. 12, 15, 23). However, in social networks, more scrutiny is required. March (1955) provides an excellent example of this when discussing interpersonal influence between two individuals, A and B. He states there is a difference "... between the influence relationship of two events (e.g. 'A votes yes,' 'B votes yes') and the relationship between two individuals (e.g. A, B)." (March, 1955, p. 435) Consequently, in measuring influence networks where the nodes are individuals, March (1955) infers (and references one of his previous works, March, 1953-54, pp. 469-470),

> ... that the appropriate model for the description of an influence relationship between two individuals is one in which the influence-related activities of the individuals are partitioned into mutually-exclusive sets such that within each set asymmetry holds between the individual agents of the activities ... (p. 436)

Consequently, exploratory analysis using a single category of events, which March (1953-54, 1955) defines as subsets of an individual's activities, can lead to an inadequate representation of the true SIN (March, 1953-54, pp. 469-470; March, 1955, p. 436). Nevertheless, as a foundational step, this research examines a single event category, i.e. a single relation, in exploratory analysis (Wasserman & Faust, 1994). The following steps comprise the method for identifying possible SIN structures, of which one or more may be selected according to theoretical or constraint criteria, and analyzed further (Klir, 1985; Shipley, 2002, p. 290; Spirtes *et al.*, 2000, pp. 124-125).

1. Identification of a set of individuals.
2. Identify event categories (potential categories could be derived from the hierarchies of Lenski (1954) and topical areas of March (1953-54)).

   For this step, an important consideration is the ability to readily measure events pertaining to influence among network members. For example, consider a clandestine network that has some voting process, where observation of the process and results is unlikely. Consequently, event proxies will generally have to be used, leading to some inaccuracies. An example of an event proxy regarding influence can be the number of meetings jointly attended, i.e. data pertaining to social closeness. (Renfro, 2001, pp. 2, 97; Watts, 2003)

3. Perform exploratory analysis using event category data. The results can be interpreted as candidate SIN structures.

   If desired, confirmatory analysis could also be accomplished by comparing a proposed 'ground truth' SIN to the exploratory analysis results (Shipley, 2002, pp. 102-103). Results from exploratory analysis may include networks with some edges not fully oriented, indicating observationally equivalent models (Glymour & Spirtes, 1988; Shipley, 2002, pp. 256-260, 265, 287-290; Spirtes *et al.*, 1993, pp. 180-183; Spirtes *et al.*, 2000, pp. 6, 59, 61, 82-87, 139-140; Verma & Pearl, 1991). In such a case, the confirmation process should account for such discrepancies.

It is arguably difficult to obtain a ground truth SIN, but it could be assumed that certain networks are reasonable proxies for 'true' social influence networks (Killworth & Bernard, 1976). For example, it may be plausible that the formal (command) hierarchy of an organization could serve as ground truth, against which network structures obtained via exploratory analysis could be validated. It is likely that such a proxy may not be entirely accurate, but for the most part it should prove adequate (Weiss & Jacobson, 1955, p. 662). This fact is seen in a study by Jacobson and Seashore (1951) involving an organization, where most individuals perceived those in a direct line of authority over them as power figures; yet, there did exist individuals who exhibited less or more power than expected given their formal position (Jacobson & Seashore, 1951, pp. 38-39). Since influence and power have been treated as synonyms, albeit in a political context, the formal organizational hierarchy could be assumed to adequately represent the ground truth social influence network (Simon, 1953, p. 501).

Results from exploratory analysis may contain a bi-directed arc between two variables, indicating the presence of a hidden variable (Shipley, 2002, pp. 256, 266-267; Spirtes *et al.*, 2000, pp. 125, 144-145; Glymour *et al.*, 2004a; Verma & Pearl, 1991). In the single relation SIN previously discussed, such a variable indicates a hidden individual in that event category. This is due to the assumption (based on the reasoning provided by March) that the association between events corresponds to influence between individuals in that particular activity (March, 1953-54, pp. 469-470; March, 1955, pp. 435-436). Consequently, revealing a hidden individual in a SIN is accomplished by simply performing exploratory analysis, and examining the results for appropriate indicators. In order to validate this approach, the event category data for an individual in the SIN will be removed, subsequent exploratory analysis conducted, and the output examined for appropriate hidden variable indicators (Shipley, 2002, pp. 266-267; Spirtes *et al.*, 2000, pp. 144-145; Glymour *et al.*, 2004a). This section has provided an overview of a method to determine SIN structures and hidden individuals within a SIN.

3.5 Conclusion

This chapter presented the approaches for the development and demonstration of three methods to address unrevealed elements in networks. The next chapter provides results of the first method.

4. Revealing Bridges in Social Networks

4.1 Introduction

This chapter provides results demonstrating the method for revealing bridge elements in social networks. The feasibility is demonstrated on both empirical and generated data sets (Jackson, Boggs, Nash & Powell, 1991).

4.2 Method

As indicated in the literature, liaisons (and by assumption bridges) may exhibit distinctive characteristics. The features can serve as potential demographic and structural covariates for a logistic regression model employed to provide insight into revealing bridges. The model could be established so that the response variable, y, equals 1 if the member is a bridge and 0 otherwise; furthermore,

$$E(y) = \pi = P(y = 1) = \frac{\exp(\mathbf{x}'\beta)}{1 + \exp(\mathbf{x}'\beta)}$$

where $\mathbf{x}$ and β would represent the covariate values and parameters, respectively (Hosmer & Lemeshow, 2000; Kleinbaum & Klein, 2002; Montgomery *et al.*, 2001).

Two standard approaches (cf. Hosmer and Lemeshow (2000), and Montgomery *et al.* (2001)) were employed to examine the bridge identification. Both approaches were implemented for the empirical data set, and one approach was employed for the generated data. The first approach involved fitting a logistic regression model (using characteristics discussed in the literature review) to the entire data set, and then examining the ability of the fitted model to identify, i.e. classify, each individual observation in the data set. This goodness of fit test constituted a bridge identification process. The goodness of fit test was performed using two cut-points. The first cut-point was the value corresponding to the maximum difference between the sensitivity and 1-specificity values of the associated Receiver Operating Characteristic (ROC) curve (Hosmer & Lemeshow, 2000; JMP 6.0.0, 2005, Nominal Logistic Regression section). The second cut-point was 0.5, which represented an uninformed assignment of a group member to either the bridge or non-bridge class (Hosmer & Lemeshow, 2000). The second identification approach involved partitioning the data set into estimation and prediction data sets. The prediction data included all individuals in groups that contained only one bridge, but more than two members. The estimation data consisted of the individuals in the remaining groups (the rationale behind the choice of the prediction data was both the consistency and convenience such a partitioning provided in performing the bridge detection method). A model was fit to the estimation data, and the fitted model was then used to classify prediction data members (Montgomery *et al.*, 2001). In a similar fashion, both a ROC curve cut-point and an uninformed cut-point were employed.

The bridge detection method was performed under the assumption that each group communicated with at least one other group and inter-group communication occurred only through bridge members. With this assumption, consider the case where a group contains a single bridge, i.e. the group's external communications occur only through that member. Suppose the group is then altered by removing the bridge and its associated ties from the group, i.e. no structural or demographic evidence of the bridge remains. If subsequent classification of each member in the altered group yields no bridges, then this may indicate the existence of an unknown bridge or bridges. Consequently, bridge detection is performed by applying the bridge identification model to groups that have had the bridge and its links removed, and surveying the resulting group member classifications. The associated claim and contribution of this research is that if groups communicate via bridges, and if bridges have demographic and structural attributes that distinguish them from non-bridges in their groups, then this method provides a feasible approach to detecting the presence of an unknown bridge or bridges in the group.

The detection approaches were similar to those of bridge identification, with both approaches applied to the empirical data and the second approach applied to the generated data. The first approach involved fitting a logistic regression model to the entire data set except those members that were the sole bridges in their groups, i.e. the data set did not contain the sole bridges or their links. Examination of the model's classification results for members in the altered single bridge groups comprises the unknown bridge detection process; based on the previously stated assumptions, where lack of an identified bridge in such groups indicates an unknown bridge or bridges. In other words, the detection method is accomplished using a goodness of fit test, i.e. comparison of member classification predicted by the model to the actual label of the member. The goodness of fit test is performed using both ROC curve and uninformed cut-points. The second detection approach partitions the data set into estimation and prediction data sets in a manner similar to the hidden bridge identification process. The exception is that the prediction data does not include the sole bridges nor their links. The model fit to the estimation data is used to classify prediction data members, using both a ROC curve cut-point and an uninformed cut-point. The same assumption holds: lack of an identified bridge in such groups indicates an unknown bridge.

One note about the detection approach is warranted. In order to demonstrate that the detection approach is feasible, the prediction data contains groups composed of non-bridge members and a single bridge. It appears logical, in most cases, that the approach would be feasible if $k > 1$ bridges were present in a group; however, in any case in which the detection approach results in none of the groups' members being classified as a bridge, the only inference that can be drawn about missing members is that there is *at least* one bridge member missing. Consequently, the prediction data used in this dissertation is appropriate for demonstrating the detection concept.

This section included rationale, assumptions and steps of approaches for revealing bridges in social networks. Application to empirical and generated data sets suggests the feasibility of the approaches, and in some instances, rather promising results as discussed in the next section.

4.3 Data Sets and Results

The above methods were implemented using both empirical and generated data. The empirical data is referred to as the Sageman terrorist data set. Sageman (2004a; 2004b) compiled open source data that contains detailed demographic and relation information for 366 individuals associated with the Global Salafi movement (Sageman, 2004b, p. 138). (The data set was provided to Clark (2005) by Sageman and will be referenced as Sageman (2004a)) The generated data consisted of 200 groups; 100 were 2-stars and 100 were 3-stars. Many group structures are possible, but in order to impose distinctive structural attributes to the members and demonstrate feasibility of identification and detection approaches, a k-star structure, $k \in \{2, 3\}$, was chosen. A k-star is a graph in which k vertices are connected to a central vertex, and no other connections exist in the graph (Frank & Strauss, 1986; Freeman, 1977; Hunter, Goodreau, & Handcock, 2008; West, 2001). While specific inter-group connections were not made, each group had a single bridge through which inter-group communication would flow under the assumptions of this research. Data set details and results of analysis are provided in the following sections.

4.3.1 Sageman Terrorist Data Set: Description

In order to apply the methods, it was necessary to define a group. For this research, a group within the Sageman (2004a) data set was defined as individuals involved in the same terrorist operation; furthermore that operation was the member's only operation in which they were involved, i.e. linking pins were not considered (Likert, 1961). Additionally, the group definition required at least two members, one of which had to be a bridge. Given this definition, the groups analyzed contained only bridges and non-bridges; therefore, while the operation may have included other members, who were involved in multiple operations, such members were not considered part of the group. This group definition simplifies the analysis, but focuses the research and its results specifically on bridges and non-bridges.

The Sageman (2004a) terrorist data set ties include acquaintance, friendship, nuclear family, relatives, teacher, religious leader and post-join ties. Post-join ties are ties formed after an individual joined the movement. The teacher and religious leader relations consisted of non-reciprocated ties; however, for this research, such ties were considered reciprocated. The remaining relations consisted of reciprocated ties; however, in certain instances some ties were not reciprocated. In these cases, it was assumed the ties should have been reciprocated, and such ties were added. Since all ties were treated as reciprocated for this analysis, the terrorist network consists of undirected links. Consequently, inter-group communication is represented by one or more undirected links between two or more groups. This yields another assumption: a link between two or more groups represents communication regardless if the original relation consisted of either non-reciprocated ties or ties that do not necessarily imply transfer of information, e.g. a tie between relatives. Furthermore, a single link was assumed to exist between two individuals even when more than one type of tie existed. Any ties from a member to itself were deleted. These assumptions and constraints produced a social network with no loops and no more than a single undirected link between any two individuals, based on the adjusted Sageman data. Note: Many of these assumptions were derived from Clark (2005).

After applying the above definitions and assumptions, the final data set consisted of seventeen groups and a total of one hundred seventy-two individuals of which fifty were bridges and one hundred twenty-two were non-bridges. Four of the seventeen groups contained more than two non-bridges, but only a single bridge. These four groups contained a total of forty-seven individuals.

The following ten covariates were examined; rationale for their use was provided in the literature review. The decision for determining the data type of each element is based on information in Clark (2005), Dillon and Goldstein (1984) and Sageman (2004a).

1. The age at which an individual joined the movement. This was treated as a continuous covariate.
2. The education level of an individual. This covariate was treated as categorical with level 0 representing less than a Master's degree and level 1 otherwise.
3. The individual's position within a group. This covariate was treated as categorical with level 0 representing a subordinate and level 1 otherwise, e.g. a position involving logistics associated with the movement.
4. The individual's family socioeconomic status was treated as a categorical variable with level 0 representing lower and middle classes, and level 1 represented an upper class status.
5. The individual's number of intra-group links; treated as a continuous covariate.
6. The individual's number of intra-group links normalized according to the group size; treated as a continuous covariate.
7. The individual's intra-group betweenness centrality; treated as a continuous covariate.
8. The individual's intra-group betweenness centrality normalized according to the number of individuals in the group; treated as a continuous covariate.
9. An individual's similarity to the other individuals in its group on the basis of age and normalized intra-group links. This continuous covariate was derived from similarity definitions/notation provided by Everitt, Landau and Leese (2001) and Gower (1971) and is equivalent to

$$S_i = \frac{\displaystyle\sum_{j=1}^{n} \frac{1}{p} \sum_{k=1}^{p} s_{ijk}}{n-1}$$

with $j \neq i$ and n is equal to the number of individuals in the group of which individual i is a member. Furthermore, p is the number of variables for which similarity is derived, which is two (age and normalized intra-group degree) for this covariate. The term s_{ijk} measures the similarity of individuals i and j for variable k, and is scored as $s_{ijk} = 1 - \frac{|x_i - x_j|}{R_k}$, where R_k is the range of variable k for individuals in the same group. The value of S_i lies in the interval [0,1], with a value of 1 indicating that individual i is exactly alike to all other group members with respect to the two attributes.

10. An individual's similarity to other individuals in its group, with respect to the joining age, family socioeconomic status and education level. This continuous covariate is defined similar to the previous covariate, except $p = 3$ and some of the attributes, e.g. family socioeconomic status and education level, are different.

Covariates 1, 2 and 4 are demographic in nature, covariates 5-8 are structural, covariate 3 is organizational (i.e. position) and the remaining two are heterophilic. From Sageman (2004a; 2004b), it appears the demographic covariates were attributes possessed by group individuals when they joined the movement; a potential exception is education level (Sageman, 2004b, pp. 103-104). This is useful because more credence can be placed on the explanatory variables contributing to the bridge or non-bridge nature of individuals; rather than the individual obtaining such attributes after fulfilling the bridge role. It is important to note that the covariates chosen were demographic or intra-group structural attributes, and not reliant upon inter-group attributes. The only instances for which covariates were assigned values related to non-group members was in the imputation of missing data. Missing link data was imputed as previously mentioned. Missing demographic data was imputed as follows (based on techniques from Clark (2005), Everitt *et al.* (2001), and Little and Rubin (1987)):

1. Joining age.

This data was calculated as the mean of the joining age of all bridge and non-bridge members in the data set of interest. Care was taken to include attribute data from only appropriate members. For example, if data was partitioned into estimation and prediction data sets, then attribute imputation for members in the estimation data set were based only on members of the estimation data set.

2. Socioeconomic status.

   If an individual did not have a socioeconomic status datum, the value assigned was the mode of the socioeconomic status of the bridges and non-bridges in the appropriate data set. The imputation process builds on the categories provided by Sageman (2004a): upper class (1), middle class (2) and lower class (3); therefore, if a member was missing a status datum, it was imputed as the mode of the status of the individuals (both bridges and non-bridges) in the appropriate data set. Subsequently, this mode was re-categorized, according to the categories in this dissertation, resulting in the final imputed status datum. Categories differing from those of Sageman (2004a) were used in this dissertation to maintain a reasonable number of explanatory variables as discussed by Hosmer and Lemeshow (2000) based on a study by Peduzzi, Concato, Kemper, Holford and Feinstein (1996).

3. Position.

   If an individual did not have a position datum, the value assigned was the mode of the position of the bridges and non-bridges in the appropriate data set. The imputation process builds on the categories of Sageman (2004a): emir (1), military committee (2), religious/fatwa committee (3), finance/logistics (4), media/propaganda (5), local chief (6) and subordinate (7); therefore, if a member was missing a position datum, it was imputed as the position mode of the position of the individuals (both bridges and non-bridges) in the appropriate data set. Subsequently, this mode was re-categorized, according to the categories in this dissertation, resulting in the final imputed position datum.

In the provided data, some individuals were attributed more than one position within the organization; however, the multiple positions never included the subordinate position, and in this study these individuals were assigned a position covariate value of 1. Additionally, these individuals' data values were ignored when the mode was calculated for imputation purposes.

4. Education level.

   If an individual did not have an education level datum, the value assigned was the mode of the education level of the bridges and non-bridges in the appropriate data set. The imputation process builds on the categories of Sageman (2004a): less than a high school graduate (1), high school graduate (2), some college (3), bachelor's degree (4), master's degree (5), and doctoral degree (6); therefore, if a member was missing an education level datum, it was imputed (using Clark's categories) as the mode of of the education level individuals (both bridges and non-bridges) in the appropriate data set. Subsequently, this mode was re-categorized, according to the categories in this dissertation, resulting in the final imputed education level datum.

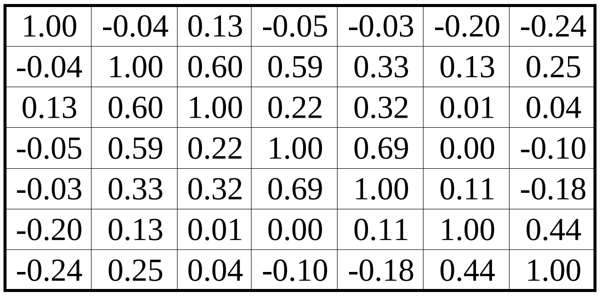
   This subsection described the Sageman data set and associated assumptions and definitions. The results of the proposed analytic approaches are provided in the next subsection.

4.3.2 Sageman Terrorist Data Set: Analysis and Results

The first approach to hidden bridge analysis incorporated the full data set of 172 individuals in seventeen groups. The seven continuous variables (ordered as Age, Intra-Group Degree, Intra-Group Normalized Degree, Intra-Group Betweenness, Intra-Group Normalized Betweenness, Similarity with respect to Age, Family

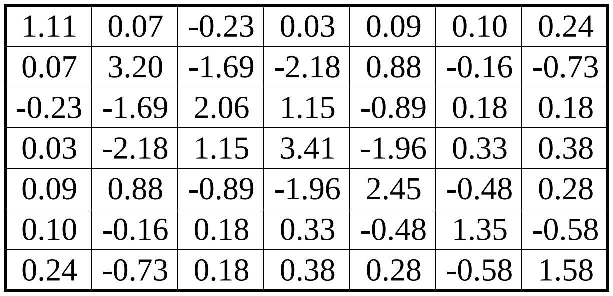
Socioeconomic Status and Education Level, and Similarity with respect to Age and Intra-Group Normalized Degree) were examined for multi-collinearity. Table 1 provides the $\mathbf{W'W}$ matrix, and Table 2 shows the $(\mathbf{W'W})^{-\mathbf{1}}$ matrix (note: the displayed entries of these and subsequent $\mathbf{W'W}$ and $\mathbf{W'W}^{-1}$ matrices are rounded values). The absolute value of the largest off-diagonal entry in $\mathbf{W'W}$ is less than 0.7, and the variance inflation factors were less than 4.0. Given these values, multi-collinearity was present, but not deemed to be a large problem (Montgomery *et al.*, 2001, pp. 334, 337).

Table 1 $\mathbf{W'W}$ Matrix for Continuous Variables in Sageman Data



1.00	-0.04	0.13	-0.05	-0.03	-0.20	-0.24
-0.04	1.00	0.60	0.59	0.33	0.13	0.25
0.13	0.60	1.00	0.22	0.32	0.01	0.04
-0.05	0.59	0.22	1.00	0.69	0.00	-0.10
-0.03	0.33	0.32	0.69	1.00	0.11	-0.18
-0.20	0.13	0.01	0.00	0.11	1.00	0.44
-0.24	0.25	0.04	-0.10	-0.18	0.44	1.00

Table 2 $(\mathbf{W'W})^{-1}$ Matrix for Continuous Variables in Sageman Data



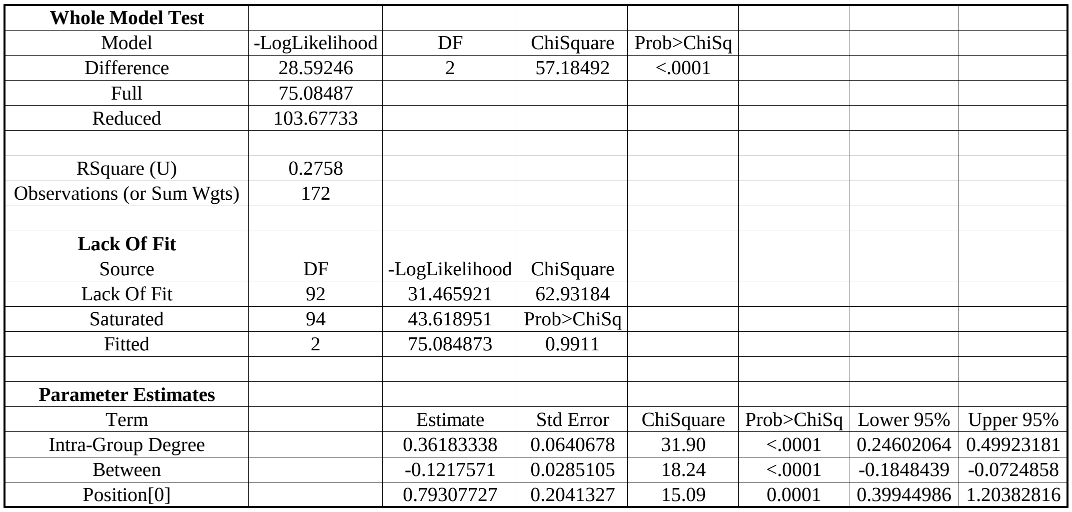
1.11	0.07	-0.23	0.03	0.09	0.10	0.24
0.07	3.20	-1.69	-2.18	0.88	-0.16	-0.73
-0.23	-1.69	2.06	1.15	-0.89	0.18	0.18
0.03	-2.18	1.15	3.41	-1.96	0.33	0.38
0.09	0.88	-0.89	-1.96	2.45	-0.48	0.28
0.10	-0.16	0.18	0.33	-0.48	1.35	-0.58
0.24	-0.73	0.18	0.38	0.28	-0.58	1.58

Using a combination of fitting all covariates and forward stepwise regression (with probability to enter = 0.25 and probability to leave = 0.1), the estimated logit, $\hat{g}(\mathbf{x})$, constructed was

$$\hat{g}(\mathbf{x}) = -0.362 Intra-GroupDegree + 0.122 Intra-GroupBetweenness - 0.793 Position$$

Table 3 provides relevant model information (Garson, n.d.; Hosmer & Lemeshow, 2000; JMP 6.0.0, 2005; Montgomery *et al.*, 2001, pp. 453-454). JMP employs categorical variable coding different from the coding listed above; therefore, 0/1 coding in this dissertation is realized as 1/-1, respectively, in JMP. Additionally, JMP reports estimates in accordance with a log odds of 0/1, so the signs of the parameter estimates are opposite those in the above logit. Finally, "RSquare(U) is the proportion of the total uncertainty that is attributed to the model fit" (JMP 6.0.0, 2005, Nominal Logistic Regression section). The sign of the degree coefficient was different from that suggested in some literature; however, the sign was reasonable from a clandestine operation perspective. Table 4 provides the identification accuracy, i.e. goodness of fit, results for the model. With the exception of bridge accuracy in the case of an uninformed cut-point, a reasonable level of accuracy was achieved demonstrating that, given the data and associated analysis, the identification approach is promising.

Table 3 Information for Model Fit to Sageman Data



Whole Model Test							
Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq			
Difference	28.59246	2	57.18492	<.0001			
Full	75.08487						
Reduced	103.67733						
RSquare (U)	0.2758						
Observations (or Sum Wgts)	172						
Lack Of Fit							
Source	DF	-LogLikelihood	ChiSquare				
Lack Of Fit	92	31.465921	62.93184				
Saturated	94	43.618951	Prob>ChiSq				
Fitted	2	75.084873	0.9911				
Parameter Estimates							
Term		Estimate	Std Error	ChiSquare	Prob>ChiSq	Lower 95%	Upper 95%
Intra-Group Degree		0.36183338	0.0640678	31.90	<.0001	0.24602064	0.49923181
Between		-0.1217571	0.0285105	18.24	<.0001	-0.1848439	-0.0724858
Position[0]		0.79307727	0.2041327	15.09	0.0001	0.39944986	1.20382816

The second approach to hidden bridge analysis employed an estimation data set of the 125 individuals in thirteen groups, each containing more than one bridge. Table 5 provides the $\mathbf{W'W}$ matrix, and Table 6 gives the $(\mathbf{W'W})^{-\mathbf{1}}$ matrix. The absolute value of the largest off-diagonal entry in $\mathbf{W'W}$ is approximately 0.7, and

Table 4 Identification Accuracy for Sageman Data Set

	ROC cut-point = 0.2396	Uninformed cut-point = 0.5
Overall Accuracy	76.74% (132 individuals)	77.33% (133 individuals)
Bridge Accuracy	76% (38 bridges)	48% (24 bridges)

the largest variance inflation factor value is approximately 4.5. Given these values, multi-collinearity was present, but not deemed to be a large problem.

Table 5 $\mathbf{W'W}$ Matrix for Continuous Variables in Sageman Estimation Data

1.00	-0.01	0.21	-0.09	-0.08	-0.20	-0.26
-0.01	1.00	0.48	0.69	0.43	0.20	0.29
0.21	0.48	1.00	0.28	0.41	0.01	-0.12
-0.09	0.69	0.28	1.00	0.71	-0.01	-0.05
-0.08	0.43	0.41	0.71	1.00	0.16	-0.12
-0.20	0.20	0.01	-0.01	0.16	1.00	0.46
-0.26	0.29	-0.12	-0.05	-0.12	0.46	1.00

Table 6 $(\mathbf{W'W})^{-1}$ Matrix for Continuous Variables in Sageman Estimation Data

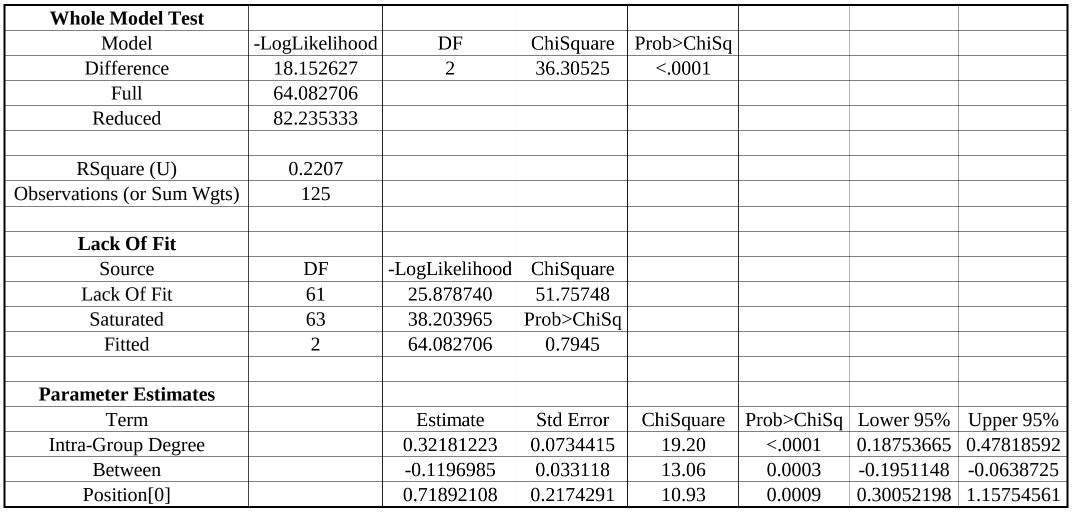
1.19	-0.24	-0.22	0.28	0.11	0.12	0.33
-0.24	3.81	-1.50	-3.02	1.01	-0.43	-1.20
-0.22	-1.50	1.87	1.18	-0.94	0.17	0.48
0.28	-3.02	1.18	4.51	-2.38	0.74	0.71
0.11	1.01	-0.94	-2.38	2.77	-0.70	0.16
0.12	-0.43	0.17	0.74	-0.70	1.48	-0.55
0.33	-1.20	0.48	0.71	0.16	-0.55	1.80

Using a combination of fitting all covariates and forward stepwise regression (with probability to enter = 0.25 and probability to leave = 0.1), the estimated logit was

$$\hat{g}(\mathbf{x}) = -0.322 Intra - GroupDegree + 0.120 Intra - GroupBetweenness - 0.719 Position$$

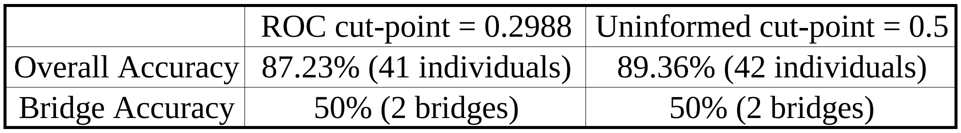
Table 7 provides relevant model information. Applying the previously fit logit and associated cut-points to the prediction data (i.e. 47 members in four groups with multiple non-bridges, but only one bridge per group) yielded identification accuracy results shown in Table 8 (Hosmer & Lemeshow, 2001, p. 186). The achieved accuracy, for the provided data and associated analysis, suggests the feasibility of the second identification approach for non-bridge members. Bridge identification results do not appear as promising; however, this may be an artifact of the small number of bridges, four, in the prediction data. Recall the data set partitioning was chosen for consistency and convenience pertaining to the bridge detection process; therefore, a different partitioning may result in greater bridge identification accuracy. For informational purposes, Table 9 provides identification accuracy of the estimation data.

Table 7 Information for Model Fit to Sageman Estimation Data



Whole Model Test								
Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq				
Difference	18.152627	2	36.30525	<.0001				
Full	64.082706							
Reduced	82.235333							
RSquare (U)	0.2207							
Observations (or Sum Wgts)	125							
Lack Of Fit								
Source	DF	-LogLikelihood	ChiSquare					
Lack Of Fit	61	25.878740	51.75748					
Saturated	63	38.203965	Prob>ChiSq					
Fitted	2	64.082706	0.7945					
Parameter Estimates								
Term		Estimate	Std Error	ChiSquare	Prob>ChiSq	Lower 95%	Upper 95%	
Intra-Group Degree		0.32181223	0.0734415	19.20	<.0001	0.18753665	0.47818592	
Between		-0.1196985	0.033118	13.06	0.0003	-0.1951148	-0.0638725	
Position[0]		0.71892108	0.2174291	10.93	0.0009	0.30052198	1.15754561	

Table 8 Identification Accuracy for Sageman Prediction Data Set



	ROC cut-point = 0.2988	Uninformed cut-point = 0.5
Overall Accuracy	87.23% (41 individuals)	89.36% (42 individuals)
Bridge Accuracy	50% (2 bridges)	50% (2 bridges)

The first approach to bridge detection required the full data set without the four bridges that were the only bridges in their groups, i.e. 168 individuals in seventeen groups. Table 10 provides the $\mathbf{W'W}$ matrix, and Table 11 gives the $(\mathbf{W'W})^{-1}$

Table 9 Identification Accuracy for Sageman Estimation Data Set

	ROC cut-point = 0.2988	Uninformed cut-point = 0.5
Overall Accuracy	73.6% (92 individuals)	71.2% (89 individuals)
Bridge Accuracy	78.26% (36 bridges)	50% (23 bridges)

matrix. The absolute value of the largest off-diagonal entry in $\mathbf{W}'\mathbf{W}$ is less than 0.7, and the variance inflation factors were less than 3.4. Given these values, multicollinearity, while present, was not deemed to be a large problem.

Table 10 $\mathbf{W}'\mathbf{W}$ for Continuous Variables in Sageman Data - No Sole Bridges

1.00	-0.05	0.14	-0.05	-0.01	-0.21	-0.24
-0.05	1.00	0.56	0.59	0.33	0.14	0.24
0.14	0.56	1.00	0.20	0.31	0.05	0.01
-0.05	0.59	0.20	1.00	0.69	-0.01	-0.08
-0.01	0.33	0.31	0.69	1.00	0.12	-0.16
-0.21	0.14	0.05	-0.01	0.12	1.00	0.44
-0.24	0.24	0.01	-0.08	-0.16	0.44	1.00

Table 11 $(\mathbf{W}'\mathbf{W})^{-1}$ for Continuous Variables in Sageman Data - No Sole Bridges

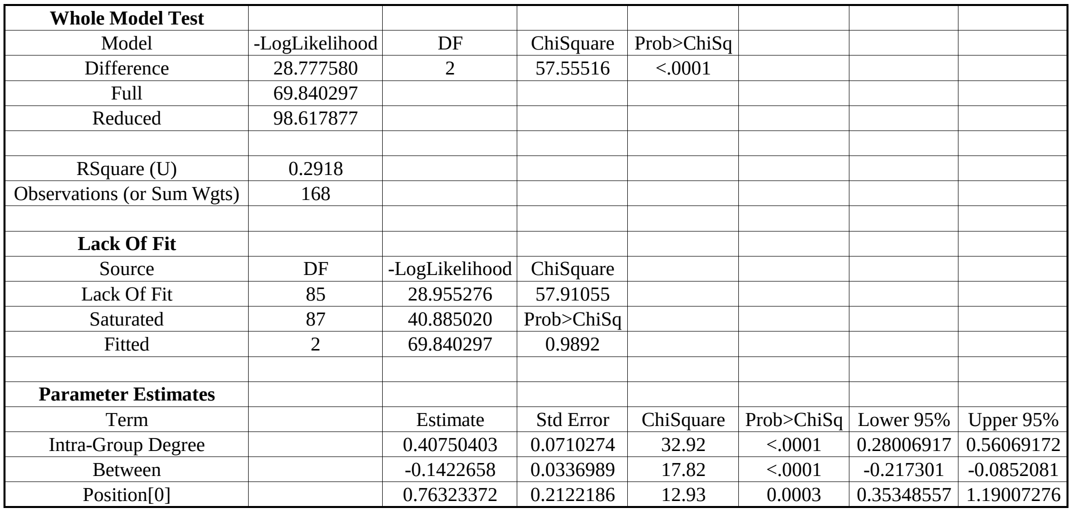
1.12	0.06	-0.22	0.06	0.04	0.14	0.21
0.06	2.96	-1.50	-2.05	0.81	-0.16	-0.66
-0.22	-1.50	1.91	1.07	-0.82	0.09	0.21
0.06	-2.05	1.07	3.38	-1.98	0.40	0.29
0.04	0.81	-0.82	-1.98	2.45	-0.51	0.27
0.14	-0.16	0.09	0.40	-0.51	1.37	-0.58
0.21	-0.66	0.21	0.29	0.27	-0.58	1.53

Using a combination of fitting all covariates and forward stepwise regression (with probability to enter = 0.25 and probability to leave = 0.1), the estimated logit was

$$\hat{g}(\mathbf{x}) = -0.408 Intra - GroupDegree + 0.142 Intra - GroupBetweenness - 0.763 Position$$

Table 12 provides relevant model information. Table 13 provides the identification accuracy results for members in the four groups altered by removal of their sole bridges and associated links. Under previous assumptions, both the ROC and uninformed cut-points gave possible indications that three of the four altered groups contained an unknown bridge, i.e. a 75% detection accuracy. In altered group C, the three non-bridges mis-identified (via the ROC cut-point) as bridges had few intra-group links, small intra-group betweenness values and positions other than a subordinate. The single mis-identification (in the altered groups) with the uninformed cut-point, was one of the mis-identifications occurring with the ROC cut-point. The position of the non-bridges appears as a plausible reason for the mis-identifications. Identification accuracy (via the ROC cut-point) of not only the altered group members, but also all other members was 77.98% (131 members), and 73.91% (34) of the bridges were accurately identified. Respective identification accuracy with the unbiased cut-point was 79.17% (133 members), and 43.48% (20 bridges). The 75% detection accuracy tentatively demonstrates the feasibility of the detection method, but choosing other groups (with possibly different structures) from which to remove bridges could yield similar or different results.

Table 12 Information for Model Fit to Sageman Data - No Sole Bridges



Whole Model Test							
Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq			
Difference	28.777580	2	57.55516	<.0001			
Full	69.840297						
Reduced	98.617877						
RSquare (U)	0.2918						
Observations (or Sum Wgts)	168						
Lack Of Fit							
Source	DF	-LogLikelihood	ChiSquare				
Lack Of Fit	85	28.955276	57.91055				
Saturated	87	40.885020	Prob>ChiSq				
Fitted	2	69.840297	0.9892				
Parameter Estimates							
Term		Estimate	Std Error	ChiSquare	Prob>ChiSq	Lower 95%	Upper 95%
Intra-Group Degree		0.40750403	0.0710274	32.92	<.0001	0.28006917	0.56069172
Between		-0.1422658	0.0336989	17.82	<.0001	-0.217301	-0.0852081
Position[0]		0.76323372	0.2122186	12.93	0.0003	0.35348557	1.19007276

The second approach to detecting unknown bridge analysis employed the same estimation data, model and cut-points as the second hidden bridge approach. Table 14 provides the identification accuracy results for members in the four altered groups.

Table 13 Identification Accuracy for Sageman Altered Groups

	Altered Group A (18 non-bridges)	Altered Group B (8 non-bridges)	Altered Group C (6 non-bridges)	Altered Group D (11 non-bridges)
ROC cutpoint = 0.2804	100%	100%	3 mis-IDs (50%)	100%
Uninformed cut-point = 0.5	100%	100%	1 mis-ID (83.33%)	100%

Under previous assumptions and with an uninformed cut-point, there are possible indications that three of the four altered groups contained an unknown bridge, i.e. a 75% detection accuracy. In altered group C, the two non-bridges mis-identified as bridges had few intra-group links, small intra-group betweenness values and positions other than a subordinate. The ROC cut-point implementation gives possible indications that two of the four altered groups contained an unknown bridge, i.e. a 50% detection accuracy. In altered group C, the three non-bridges mis-identified as bridges had few intra-group links, small intra-group betweenness values and positions other than a subordinate; while in altered group A, the mis-identified member had a position other than a subordinate. The 75% detection accuracy, in the case of the uninformed cut-point, tentatively demonstrates the feasibility of the detection method; however, ROC cut-point results are not as convincing.

Table 14 Identification Accuracy for Sageman Altered Groups (Prediction Data)

	Altered Group A (18 non-bridges)	Altered Group B (8 non-bridges)	Altered Group C (6 non-bridges)	Altered Group D (11 non-bridges)
ROC cutpoint = 0.2988	1 mis-ID (94.44%)	100%	3 mis-IDs (50%)	100%
Uninformed cut-point = 0.5	100%	100%	2 mis-IDs (66.67%)	100%

For the Sageman data and associated analysis, the identification and detection approaches showed some promise in an operational setting, with a few exceptions. Data containing more bridges was desirable and was a motivation for generating test data; however, prior to providing those results an excursion into the problem of inserting detected nodes is presented.

In groups with an inferred unknown node, the question of where to insert the unknown node remains. It is assumed, for now, that only one unknown node per group exists. Since the node is a bridge, $\pi(\mathbf{x}) = 1$; however, this is only realized in the limit since the logistic function is $\pi(\mathbf{x}) = \frac{1}{1+e^{-g(\mathbf{x})}}$ (Hosmer & Lemeshow, 2000, p. 32; Kleinbaum & Klein, 2002, pp. 5-6; Montgomery *et al.*, 2001, pp. 445, 449). Despite this asymptotic constraint, two approaches to node insertion are suggested. The simplest tactic is the following heuristic: Connect the inferred bridge to any node with which the similarity of demographic attributes is 0.5 or greater. The next tactic is to derive, possibly multiple, structures by minimizing $-\hat{g}(\mathbf{x})$ subject to feasibility constraints imposed by $\hat{g}(\mathbf{x})$ and the extant structure; thus maximizing $\hat{\pi}$ (Hosmer & Lemeshow, 2000). The following analysis examines each approach as applied to altered Group B in the second bridge detection approach to the Sageman data.

The heuristic approach assigns the inferred node, $x_{unknown}$, the average age of members, and the mode for education level, family socioeconomic status and position. Based on the eight non-bridges in altered Group B, the imputed demographic attributes of $x_{unknown}$ are age joined = 25, education level = 0, family socioeconomic status = 0, and position = 0. The resulting similarities of $x_{unknown}$ are greater than 0.5 indicating the unknown node, i.e. bridge, is connected to all other members. The resulting structure is equivalent to the ground truth structure; however, the true demographic attributes of $x_{unknown}$ were age joined = 25, education level = 1, family socioeconomic status = 0, and position = 1. Consequently, the similarity (cf. previously provided definition) between the imputed and true demographic attributes of $x_{unknown}$ is 0.5.

The second approach is as follows. It is desired to minimize

$$-\hat{g}(\mathbf{x}) = 0.322 Intra-GroupDegree - 0.120 Intra-GroupBetweenness + 0.719 Position$$

The minimum of $-g(\mathbf{x})$ occurs when Intra-Group Degree = 0, Intra-Group Betweenness = max group betweenness, and Position = -1 (in JMP coding; 1 in the original coding); however, according to centrality definitions, it is not possible for

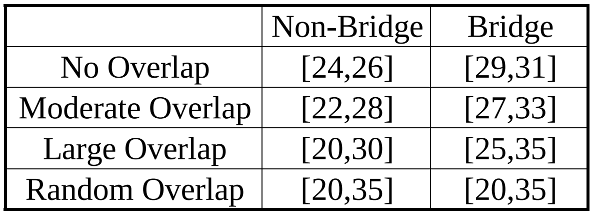
Intra-Group Degree = 0 and Intra-Group Betweenness > 0 (e.g. Freeman, 1977, pp. 36-37). Consequently, it is necessary to evaluate the structure according to degree and betweenness parametrics. The structure of the extant nodes is known, in this case, and it is a clique; consequently, the betweenness of the unknown node will always be zero, but its degree could vary from 0 to 8. Given these constraints, the minimum of $-\hat{g}(\mathbf{x})$ occurs when Intra-Group Degree = 0, Intra-Group Betweenness = 0 and Position = -1 resulting in $-\hat{g}(\mathbf{x}) = -0.719$; therefore, maximizing $\hat{\pi} = 0.672$. The true intra-group degree of the bridge was 8 not 0, but the remaining covariate values matched; consequently, the resulting structure is not equivalent to the ground truth structure.

The results of the analysis on the Sageman data give some demonstration of the feasibility of the developed method for revealing bridge elements in real world, operational social networks. Additionally, two approaches were provided for inserting detected bridges into a group. More results for the revelation method were desired; therefore, data was generated and analyzed. Results are provided in the following section.

4.3.3 Generated Data Set: Analysis and Results

In order to further examine the feasibility of the identification and detection methods, bridge and non-bridge attribute data were generated for 700 members in 100 2-stars and 100 3-stars (These members/groups were generated once, but the group member attributes, as well as, the assignment of the groups to estimation and prediction data were re-accomplished for each of the four cases discussed later in this section). Each group was assigned a single bridge, which permits demonstration of the detection method. Furthermore, the groups were not connected to form a network, *per se*; however, it is assumed each group communicates with at least one other group, and that inter-group communication occurs only via bridges. Since the detection method does not rely upon *inter*-group links, the method can still be reliably performed using the created groups, i.e. without the formal network structure. An array of demographic and structural attributes can be associated with bridges and non-bridges; however, age and intra-group degree centrality were chosen for this test. The assumption was that bridges would typically be older and have more contacts than non-bridges; however, other assumptions could easily be implemented. Additionally, even though the age and degree-centrality values were only permitted to be integers, they were treated as continuous, not ordinal, variables in the analysis (Dillon & Goldstein, 1984, pp. 2-3). The analysis approach was to increase the variability of the age and intra-group degree centrality values so that the distinction between bridges and non-bridges decreased. For each amount of variability, the detection method was implemented and results analyzed. Four cases were examined: No overlap, moderate (i.e. small) overlap, large overlap and approximately random assignment of attributes (permitting up to 100% overlap). The amount of overlap was determined by sampling a random age and intra-group degree centrality value from two different intervals, i.e. an interval for the bridge and an interval for the non-bridges. The age intervals are given in Table 15.

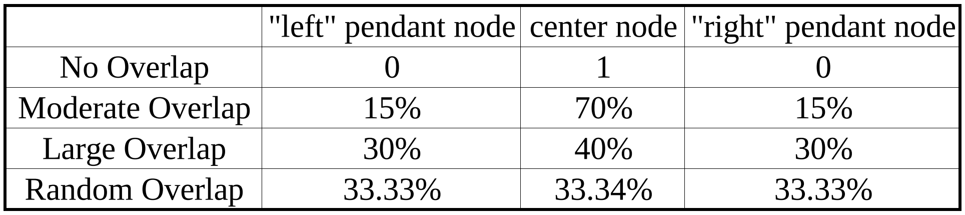
Table 15 Age Intervals for Members in Generated Groups



	Non-Bridge	Bridge
No Overlap	[24,26]	[29,31]
Moderate Overlap	[22,28]	[27,33]
Large Overlap	[20,30]	[25,35]
Random Overlap	[20,35]	[20,35]

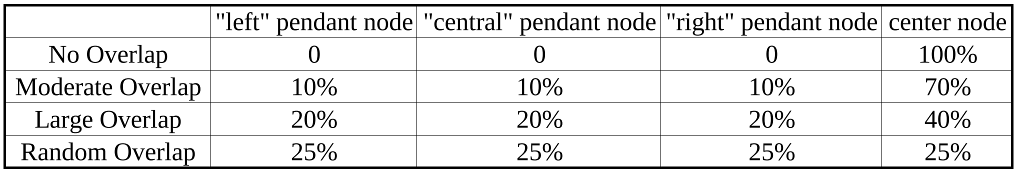
The probabilities for assigning a particular node in the 2- and 3-stars as the bridge (there is only one bridge per group) are provided in Tables 16 and 17, respectively. This assignment process constituted the generation of the degree centrality attribute of the group members. In the case of no overlap, bridges are always the center nodes, i.e. 2(3)-degree nodes, and non-bridges are always pendant, i.e. 1-degree, nodes. In the moderate and large overlap cases, the center node in the k-star is the bridge, with approximate probabilities 0.7 and 0.4, respectively. The probabilities in the tables are approximate (except for the no overlap case) due to the chosen partitioning and precision issues.

Table 16 Node/Bridge Assignment Probabilities for Generated 2-stars



	"left" pendant node	center node	"right" pendant node
No Overlap	0	1	0
Moderate Overlap	15%	70%	15%
Large Overlap	30%	40%	30%
Random Overlap	33.33%	33.34%	33.33%

Table 17 Node/Bridge Assignment Probabilities for Generated 3-stars



	"left" pendant node	"central" pendant node	"right" pendant node	center node
No Overlap	0	0	0	100%
Moderate Overlap	10%	10%	10%	70%
Large Overlap	20%	20%	20%	40%
Random Overlap	25%	25%	25%	25%

As previously mentioned, the approach for identification and detection involved randomly splitting the generated groups into estimation and prediction data. A 75% (estimation) and 25% (prediction) split of the groups was implemented. In the assignment of bridge nodes, the attribution of age to group members and data set partitioning, the random number generator seed was not tracked (more than one software was used to generate random numbers); furthermore, the order in which these tasks were performed may not have been the same for every case. There were also instances of a task/random number being re-accomplished/regenerated, while other tasks/random numbers were not. Consequently, it is possible some cases and tasks employed the same seeds resulting in group/member correlation among the cases; however, this potential correlation was ignored in the analysis.

It is important to note that structural information can be used without the regression model, in certain cases, to detect the existence of unknown members. For example, if the group structure is known to be a 2-star or 3-star, then any 2 member group implies a missing member. This is addressed in the analysis by only removing the 3-star bridges. Another example involves a group with three members, each having intra-group degree centrality 0. This also implies, in the case where groups are either 2-stars or 3-stars, a missing member. Consequently, if there is little to no variability and overlap for the structural regressor and the group structure is constrained, then detection of bridges can, in some cases, be ascertained by structural information alone. These situations are somewhat related to the topic of reconstructing a group/network from observing partial information of the group/network, which is discussed in Chapter 5.

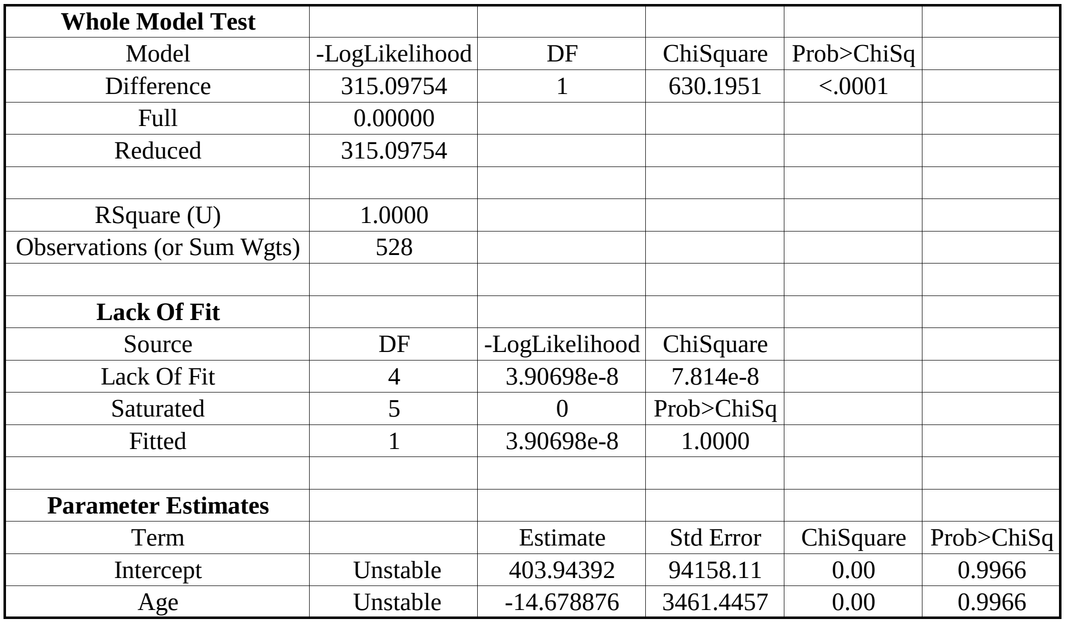
In the no overlap case, collinearity, $\rho = 0.88$, existed between age and intra-group degree centrality for the estimation data. The estimation data consisted of 150 randomly selected groups (78 3-stars and 72 2-stars) containing 528 members of which 150 were bridges. Age was chosen as the single covariate and the estimated logit, $\hat{g}(\mathbf{x})$, was

$$\hat{g}(\mathbf{x}) = -403.94 + 14.68 age$$

Table 18 provides the model information. Only models containing main effects without covariate transformations were examined. The prediction data for the identification method contained the remaining 172 members in 50 groups of which 22 were 3-stars and 28 were 2-stars. All prediction data members were correctly classified using the fitted model for both the uninformed and ROC cut-points, 0.5 and 0.99, respectively. A ROC cut-point of 1.0 was identified by the statistical software; however, for algorithmic purposes, the cut-point was treated as 0.99.

Analysis of bridge detection, for the no overlap case, involved removing all 3-star bridges from the prediction data. There were 22 such single-bridges, leaving 150 members in the prediction data. The model fit to the estimation data was used, and the goodness of fit assessment performed on the prediction data, using an unbiased cut-point of 0.5 and a ROC cut-point of 0.99, resulted in 100% accuracy in classification for both cut-points. Under previous assumptions, the 22 unknown bridges are inferred, which is understandable since no overlap exists. As previously

Table 18 Information for Model Fit to Generated Estimation Data - No Overlap



Whole Model Test					
Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq	
Difference	315.09754	1	630.1951	<.0001	
Full	0.00000				
Reduced	315.09754				
RSquare (U)	1.0000				
Observations (or Sum Wgts)	528				
Lack Of Fit					
Source	DF	-LogLikelihood	ChiSquare		
Lack Of Fit	4	3.90698e-8	7.814e-8		
Saturated	5	0	Prob>ChiSq		
Fitted	1	3.90698e-8	1.0000		
Parameter Estimates					
Term		Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	Unstable	403.94392	94158.11	0.00	0.9966
Age	Unstable	-14.678876	3461.4457	0.00	0.9966

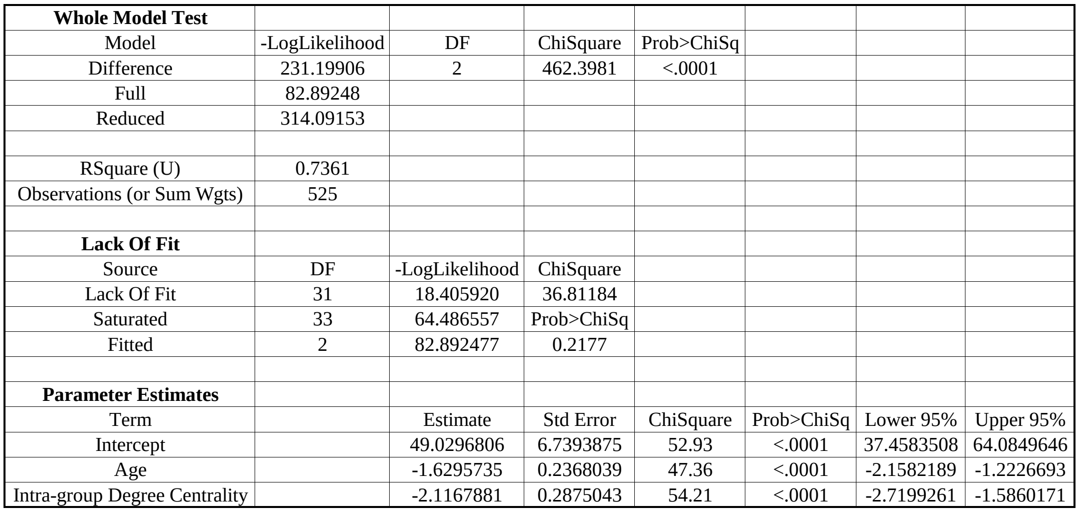
mentioned, these bridges could also be inferred by observing only the structure of the groups.

In the case of moderate overlap, the correlation coefficient between age and intra-group degree centrality for the estimation data was $\rho = 0.47$. The estimation data consisted of 150 randomly selected groups (75 3-stars and 75 2-stars) containing 525 members. The estimated logit was

$$\hat{g}(\mathbf{x}) = -49.03 + 1.63 Age + 2.12 Intra - GroupDegree$$

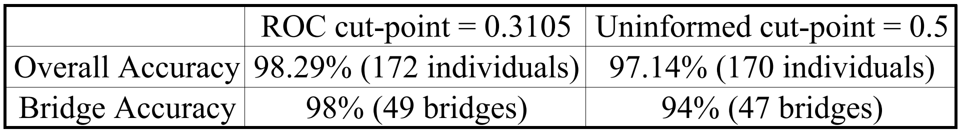
Table 19 provides the model information. Only models containing main effects without covariate transformations were examined, even though the associated lack of fit p-value was 0.2177. The prediction data contained the remaining 175 members in 50 groups of which 25 were 3-stars and 25 were 2-stars. Table 20 provides the results for the identification method, i.e. prediction data member classification accuracy. In both of the cut-point cases, the bridges misclassified as non-bridges had an age of 27 and a degree centrality of either 1 or 2. Likewise, the non-bridges classified as bridges had an age of 28 and a degree centrality of 2 or 3. The identification method is effective when a moderate overlap exists between the attributes of bridges and non-bridges.

Table 19 Information for Model Fit to Generated Estimation Data - Moderate Overlap



Whole Model Test							
Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq			
Difference	231.19906	2	462.3981	<.0001			
Full	82.89248						
Reduced	314.09153						
RSquare (U)	0.7361						
Observations (or Sum Wgts)	525						
Lack Of Fit							
Source	DF	-LogLikelihood	ChiSquare				
Lack Of Fit	31	18.405920	36.81184				
Saturated	33	64.486557	Prob>ChiSq				
Fitted	2	82.892477	0.2177				
Parameter Estimates							
Term		Estimate	Std Error	ChiSquare	Prob>ChiSq	Lower 95%	Upper 95%
Intercept		49.0296806	6.7393875	52.93	<.0001	37.4583508	64.0849646
Age		-1.6295735	0.2368039	47.36	<.0001	-2.1582189	-1.2226693
Intra-group Degree Centrality		-2.1167881	0.2875043	54.21	<.0001	-2.7199261	-1.5860171

Table 20 Identification Accuracy for Generated Groups with Moderate Overlap



	ROC cut-point = 0.3105	Uninformed cut-point = 0.5
Overall Accuracy	98.29% (172 individuals)	97.14% (170 individuals)
Bridge Accuracy	98% (49 bridges)	94% (47 bridges)

Analysis of bridge detection involved removing the 25 3-star bridges from the identification method's prediction data, appropriately modifying the degree of the remaining 3-star members, and applying the fitted model of the moderate overlap estimation data to the new prediction data, i.e. the remaining 150 members. It should be noted that pendant nodes were assigned as bridges in 5 of the 25 3-stars; therefore, removal of those bridges resulted in an 'ambiguous' structure that appeared to be a 2-star, i.e. structural inspection alone would not indicate if there was a missing group member. The detection results for the 150 members in the detection method prediction data are provided in Table 21. The detection of unknown bridges is promising for both cut-points, even in the case of ambiguous 2-stars; furthermore, only the uninformed cut-point incorrectly detected the presence of unknown bridges, and there were only 2 such false positives. Identification accuracy (via the ROC cut-point) of not only the altered group members, but also all other members was 98.67% (148 members), and 100% (25) of the bridges were accurately identified. Respective identification accuracy with the unbiased cut-point was 97.33% (146 members), and 92% (23 bridges). The detection accuracy provides a demonstration of the feasibility of the detection method in the case of moderate overlap between the assigned attributes of bridges and non-bridges in the 2- and 3-star group structures.

Table 21 Detection Accuracy for Generated Groups with Moderate Overlap

	ROC cut-point = 0.3105	Uninformed cut-point = 0.5
Correctly Inferred Unknown Bridges in 3-stars	96% (24 bridges)	96% (24 bridges)
Correctly Inferred Unknown Bridges in Ambiguous 2-stars	80% (4 bridges)	80% (4 bridges)
Incorrectly Inferred Unknown Bridges in True 2-stars	0% (0 bridges)	8% (2 bridges)

In the case of large overlap, the correlation coefficient between age and intra-group degree centrality for the estimation data was $\rho = 0.11$. The estimation data consisted of 150 randomly selected groups (78 3-stars and 72 2-stars) containing 528 members. Since the fit using only models containing main effects without covariate transformations was poor (Probability $> \chi^2 = 0.0562$), other variations were examined. The estimated logit was

$$\hat{g}(\mathbf{x}) = -1.39 + 1.52 \log(Intra - GroupDegree)$$

The correlation coefficient between age and the natural logarithm of intra-group degree centrality was $\rho = 0.13$. Table 22 provides the model information. The prediction data contained the remaining 172 members in 50 groups of which 22 were 3-stars and 28 were 2-stars. Table 23 provides the results for the identification method, i.e. prediction data member classification accuracy. As anticipated, the identification method is less promising when a large overlap exists between the attributes of bridges and non-bridges since there is less distinction.

Table 22 Information for Model Fit to Generated Estimation Data - Large Overlap

Whole Model Test							
Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq			
Difference	23.93455	1	47.86909	<.0001			
Full	291.16299						
Reduced	315.09754						
RSquare (U)	0.0760						
Observations (or Sum Wgts)	528						
Lack Of Fit							
Source	DF	-LogLikelihood	ChiSquare				
Lack Of Fit	1	0.62031	1.240625				
Saturated	2	290.54268	Prob>ChiSq				
Fitted	1	291.16299	0.2654				
Parameter Estimates							
Term		Estimate	Std Error	ChiSquare	Prob>ChiSq	Lower 95%	Upper 95%
Intercept		1.38862339	0.1268538	119.83	<.0001	1.1457102	1.64365131
Log(Intra-group Degree Centrality)		-1.5220611	0.2226925	46.71	<.0001	-1.9631274	-1.0888662

Table 23 Identification Accuracy for Generated Groups with Large Overlap

	ROC cut-point = 0.4174	Uninformed cut-point = 0.5
Overall Accuracy	61.63% (106 individuals)	61.63% (106 individuals)
Bridge Accuracy	6% (3 bridges)	6% (3 bridges)

Analysis of bridge detection involved removing the 22 3-star bridges from the identification method's prediction data, appropriately modifying the degree of the remaining 3-star members, and applying the fitted model of the large overlap estimation data to the new prediction data, i.e. the remaining 150 members. Of the 22 3-stars, 19 had an 'ambiguous' structure. The detection results for the 150 members in the detection method prediction data are provided in Table 24. The detection results are not so promising, because even though all unknown bridges were correctly inferred, all 28 2-star bridges were mis-identified thus inferring 28 unknown bridges. Identification accuracy (via both cut-points) of not only the altered group members, but also all other members was 81.33% (122 members), but 0% (0) of the bridges were accurately identified.

Given the results for the large overlap case, the previously discounted model (without the logarithmic transformation) was employed and results compared with the model in Table 22. The secondary model information is given in Table 25, and

Table 24   Detection Accuracy for Generated Groups with Large Overlap

	ROC cut-point = 0.4174	Uninformed cut-point = 0.5
Correctly Inferred Unknown Bridges in 3-stars	100% (22 bridges)	100% (22 bridges)
Correctly Inferred Unknown Bridges in Ambiguous 2-stars	100% (19 bridges)	100% (19 bridges)
Incorrectly Inferred Unknown Bridges in True 2-stars	100% (28 bridges)	100% (28 bridges)

the estimated logit was

$$\hat{g}(\mathbf{x}) = -2.2 + 0.85 Intra - GroupDegree$$

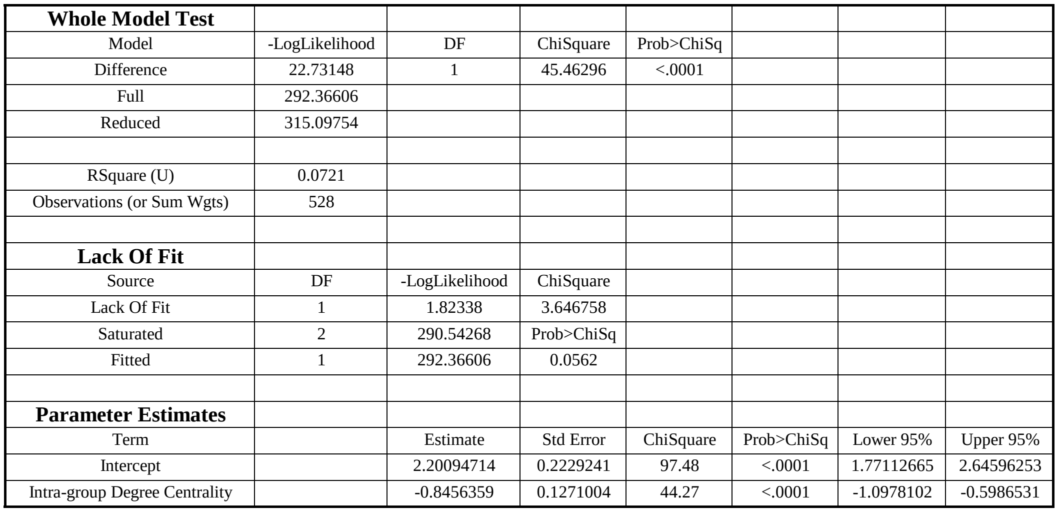
Table 26 provides the identification method results. The results are similar to the original model results, except the ROC cut-point case has less overall accuracy, but higher bridge accuracy. The secondary model detection results are provided in Table 27. The results are somewhat different than the first model, but are also not promising. While the ROC cut-point approach did not incorrectly infer bridges in the true 2-stars, there were still many misclassifications (40.48%) in the 2-stars. Identification accuracy (via the ROC cut-point) of not only the altered group members, but also all other members was 64.67% (97 members), and 39.29% (11) of the bridges were accurately identified. Respective identification accuracy with the unbiased cut-point was 81.33% (122 members), and 0% (0 bridges). For the data and attributes imposed in the large overlap case, the detection method accuracy (for either model) is less than desirable, either due to straightforward inaccuracies or false-positives.

In the random overlap case, the correlation coefficient between age and intra-group degree centrality for the estimation data, was $\rho = -0.06$. The estimation data consisted of 150 randomly selected groups (70 3-stars and 80 2-stars) containing 520 members. The estimated logit was

$$\hat{g}(\mathbf{x}) = -0.85 + 0.01 Age - 0.16 Intra - GroupDegree$$

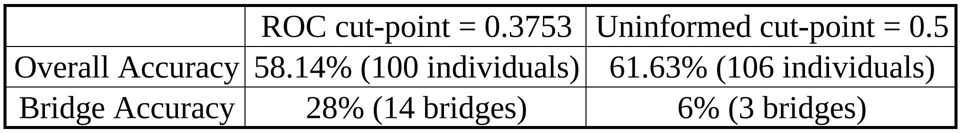
Table 28 provides the model information. The prediction data contained the remaining 180 members in 50 groups of which 30 were 3-stars and 20 were 2-stars. Table

Table 25 Information for Secondary Model Fit to Generated Estimation Data - Large Overlap



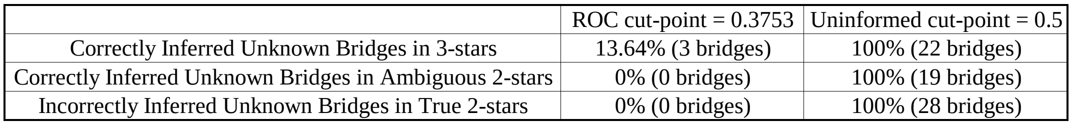
Whole Model Test							
Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq			
Difference	22.73148	1	45.46296	<.0001			
Full	292.36606						
Reduced	315.09754						
RSquare (U)	0.0721						
Observations (or Sum Wgts)	528						
Lack Of Fit							
Source	DF	-LogLikelihood	ChiSquare				
Lack Of Fit	1	1.82338	3.646758				
Saturated	2	290.54268	Prob>ChiSq				
Fitted	1	292.36606	0.0562				
Parameter Estimates							
Term		Estimate	Std Error	ChiSquare	Prob>ChiSq	Lower 95%	Upper 95%
Intercept		2.20094714	0.2229241	97.48	<.0001	1.77112665	2.64596253
Intra-group Degree Centrality		-0.8456359	0.1271004	44.27	<.0001	-1.0978102	-0.5986531

Table 26 Identification Accuracy for Generated Groups with Large Overlap - Secondary Model



	ROC cut-point = 0.3753	Uninformed cut-point = 0.5
Overall Accuracy	58.14% (100 individuals)	61.63% (106 individuals)
Bridge Accuracy	28% (14 bridges)	6% (3 bridges)

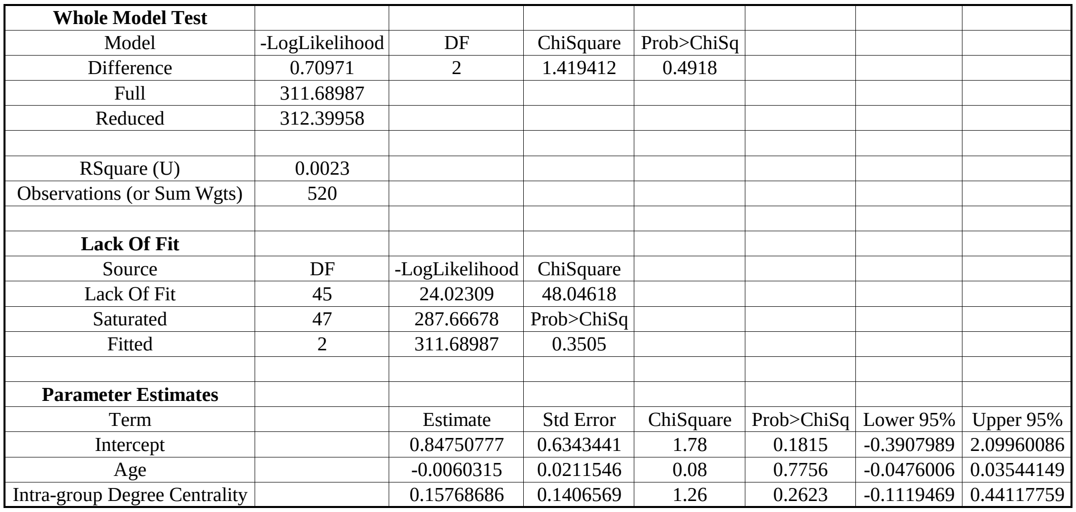
Table 27 Detection Accuracy for Generated Groups with Large Overlap - Secondary Model



	ROC cut-point = 0.3753	Uninformed cut-point = 0.5
Correctly Inferred Unknown Bridges in 3-stars	13.64% (3 bridges)	100% (22 bridges)
Correctly Inferred Unknown Bridges in Ambiguous 2-stars	0% (0 bridges)	100% (19 bridges)
Incorrectly Inferred Unknown Bridges in True 2-stars	0% (0 bridges)	100% (28 bridges)

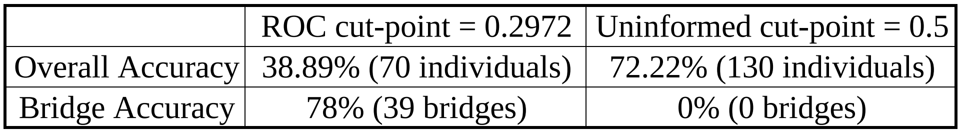
29 provides the results for the identification method, i.e. prediction data member classification accuracy. As expected, the identification method is less promising than when the attributes of bridges and non-bridges are moderately overlapped.

Table 28 Information for Model Fit to Generated Estimation Data - Random Overlap



Whole Model Test							
Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq			
Difference	0.70971	2	1.419412	0.4918			
Full	311.68987						
Reduced	312.39958						
RSquare (U)	0.0023						
Observations (or Sum Wgts)	520						
Lack Of Fit							
Source	DF	-LogLikelihood	ChiSquare				
Lack Of Fit	45	24.02309	48.04618				
Saturated	47	287.66678	Prob>ChiSq				
Fitted	2	311.68987	0.3505				
Parameter Estimates							
Term		Estimate	Std Error	ChiSquare	Prob>ChiSq	Lower 95%	Upper 95%
Intercept		0.84750777	0.6343441	1.78	0.1815	-0.3907989	2.09960086
Age		-0.0060315	0.0211546	0.08	0.7756	-0.0476006	0.03544149
Intra-group Degree Centrality		0.15768686	0.1406569	1.26	0.2623	-0.1119469	0.44117759

Table 29 Identification Accuracy for Generated Groups with Random Overlap



	ROC cut-point = 0.2972	Uninformed cut-point = 0.5
Overall Accuracy	38.89% (70 individuals)	72.22% (130 individuals)
Bridge Accuracy	78% (39 bridges)	0% (0 bridges)

Analysis of bridge detection involved removing the 30 3-star bridges from the identification method's prediction data, appropriately modifying the degree of the remaining 3-star members, and applying the fitted model of the random overlap estimation data to the new prediction data, i.e. the remaining 150 members. Of the 30 3-stars, 23 had an 'ambiguous' structure. The detection results for the 150 members in the detection method prediction data are provided in Table 30. As expected, the detection method is not consistently capable of a high detection accuracy and a low false-positive rate, when there are not discriminatory features between bridges and non-bridges. Regarding the ROC cut-point approach, there were 36 (of 60) misclassified 2-star members, but none of those incorrectly implied the existence of unknown bridges. Identification accuracy (via the ROC cut-point) of not only the altered group members, but also all other members was 26% (39 members), and 80% (16) of the bridges were accurately identified. Respective identification accuracy with the unbiased cut-point was 86.67% (130 members), and 0% (0 bridges).

Table 30 Detection Accuracy for Generated Groups with Random Overlap

	ROC cut-point = 0.2972	Uninformed cut-point = 0.5
Correctly Inferred Unknown Bridges in 3-stars	0% (0 bridges)	100% (30 bridges)
Correctly Inferred Unknown Bridges in Ambiguous 2-stars	0% (0 bridges)	100% (23 bridges)
Incorrectly Inferred Unknown Bridges in True 2-stars	0% (0 bridges)	100% (20 bridges)

Application of the proposed approaches to the generated data suggests feasibility for revealing bridge elements; especially in situations involving much dissimilarity between bridge and non-bridge members. The next section provides overall conclusions regarding the developed methods.

4.4 Conclusion

This chapter provided results demonstrating the feasibility of the developed method for revealing bridge elements in social networks. The feasibility was demonstrated on both empirical and generated data sets. Evident from the analysis, is the fact that in order for the method to provide reasonable results (i.e. high classification accuracy and/or low false positive rates) there must be attributes that distinguish a bridge from a non-bridge. While this analysis focused on social networks, the method should be feasible on other network types; assuming there are distinct connection nodes. In fact, data from a physical system network would likely be ‘cleaner’, i.e. contain less variation; consequently, improved classification results would be expected. As previously mentioned, there may exist conditions under which observations of partial network structures can result in the revelation of other network elements.

5. Network Reconstruction

5.1 Introduction

This chapter provides a new framework for reconstructing a network from its subnetworks, a general formula for reconstruction accuracy within the framework and a demonstration of the results of the associated analysis. The framework developed in this dissertation is a natural extension of graph reconstruction with applied probability, but was not found in the reviewed literature. This framework provides an approach to assessing possible network structures; furthermore, insight into potential mis-identification can be derived from the associated analysis.

5.2 Network Reconstruction Framework

The framework developed in this dissertation extends graph (i.e. network) reconstruction by adding the possibility of repeat observations. The specific problem statement associated with the framework follows: Given a n node network with unknown structure represented by an unlabeled graph, G_u, what is the duration required to reconstruct, if possible, G_u, and the accuracy of the reconstruction as time progresses (and hence the number of observations increases) under the following conditions (Bollobás, 2001, p. 42; Erdős & Rényi, 1960, p. 20; Harary, 1964; Harary & Plantholt, 1985; Myrvold, 1988, 1990)

1. There are only enough resources to observe $n-1$ nodes with their associated links in a single time step.
2. Repeat observations of a subnetwork are possible, i.e. it is not guaranteed that a subnetwork (subgraph) will be observed once and only once. Consequently, it is possible that a subgraph may be observed frequently, or not at all.

This framework and the associated definition of accuracy are related to three research areas:

1. Ally and adversary reconstruction numbers; also referred to as existential and universal reconstruction numbers (Baldwin, 2003; Harary & Plantholt, 1985; Hemaspaandra, Hemaspaandra, Radziszowski, & Tripathi, 2007; Myrvold, 1988, 1990, 1992). Essentially, the reconstruction numbers address the question of the required amount of information (i.e. duration) to reconstruct a graph (Bondy & Hemminger, 1977). Manvel (1969) proposed the question regarding the number of any (i.e. randomly selected) vertex deleted subgraphs required for reconstruction (J. Bondy, personal communication, February 5, 2008; Bondy & Hemminger, 1977).

2. The number of non-isomorphic graphs that can be reconstructed from an incomplete set of vertex deleted subgraphs, which is interwined with the previous research area (Bryant, 1971; Hemaspaandra *et al.*, 2007). The legitimate subdeck and subdeck checking problems, as given in Hemaspaandra *et al.* (2007), are closely related to this research problem. In order to understand the two (sub)problems, it is necessary to explain the legitimate vertex deck and vertex deck checking problems. In the legitimate vertex deck problem, the entire set of vertex-deleted subgraphs, i.e. a deck, is provided and one attempts to determine if the deck is legitimate, i.e. is there some graph that could have generated the deck (Bondy & Hemminger, 1977; Harary, 1969; Hemaspaandra *et al.*, 2007; Nash-Williams, 1978). Hemaspaandra *et al.* (2007) state that in the vertex deck checking problem both a deck and a graph are provided, and the question is whether the deck could have been generated by the given graph; consequently, the legitimate subdeck and subdeck checking problems are analogous to the legitimate vertex deck and vertex deck checking problems, except partial decks, i.e. subdecks, are provided instead of entire decks. Hemaspaandra *et al.* (2007) studied the computational complexity of (sub)deck problems, to include reconstructing graphs from subdecks. An example, given by Hemaspaandra *et al.* (2007), of reconstruction from subdecks is that of Harary and

Palmer (1966), who researched construction of trees from maximal subtrees. One final point is that the set reconstruction conjecture can be considered a special case of subdeck reconstruction, assuming the deck contains at least two isomorphic subgraphs.

3. The identification problem in reconstructability analysis. Similarities include attempting to identify an overall unknown system from subsystems, and capturing the behavior of the generative system via conditional probabilities (Klir, 1985, pp. 15, 114-115). There is, however, a key difference between this framework and the identification problem; specifically, the identification problem assumes the variables are labeled (Klir, 1985, p. 38). This is not the case for this reconstruction framework since the network is represented by an unlabeled graph.

This effort complements the above research areas by providing a definition and an application of network reconstruction accuracy in a probabilistic fashion. Additionally, reconstruction duration is addressed.

Prior to developing definitions, notations and formulas for the graph/network reconstruction framework with repeat observations, the analog is provided for traditional graph reconstruction.

5.2.1 Framework for Traditional Graph Reconstruction

Traditional graph reconstruction examines the reconstructability of graphs, and the number of vertex deleted subgraphs required to uniquely, up to isomorphism, reconstruct a graph (Bondy, 1991; Harary & Plantholt, 1985; Lauri, 1987, 1992; Myrvold, 1988). In this dissertation, it is assumed G_u is reconstructable; therefore, the focus is on the duration and/or accuracy of reconstructing G_u from observations of vertex deleted subgraphs. In such a context, accuracy is defined as the probability of reconstructing G_u given the current and previous subgraph observations. This definition is conceptually similar to a function describing the behavior of a generative system as provided in Klir (1985). Additionally, duration is defined as the number of observations required to reconstruct G_u. Since vertex deleted subgraphs are observed once and only once in traditional graph reconstruction, the shortest and longest reconstruction durations are equivalent to the ally and adversary reconstruction numbers (RNs), respectively (assuming the reconstruction conjecture is true) (Hemaspaandra *et al.*, 2007; Myrvold, 1988).

The following constraints, definitions and notation comprise the framework applied to traditional graph reconstruction.

1. G_u: An undirected, reconstructable graph from which vertex deleted subgraphs are observed. G_u is unknown to the observer, and represents an unknown network structure.
2. $|G_u|$ is the number of vertices of G_u; also referred to as the order of G_u (Myrvold, 1990). $|G_u|$ is known to the observer.
3. S_u is the set of vertex deleted subgraphs of G_u, and $|S_u| = |G_u|$. Since the vertex deleted subgraphs in S_u may not be unique, S_u can be considered a multiset (Blizard, 1989; Baldwin, 2003; Harary, 1964; Hemaspaandra *et al.*, 2007; Kelly, 1957).
4. Each vertex deleted subgraph may be observed only once; consequently, the number of observations is in $[1, \ldots, |G_u|]$. Note: If non-isomorphic vertex deleted subgraphs were observed instead of vertex deleted subgraphs, then this framework could address set reconstruction.
5. G_i is the ith isomorphism class of a graph with $|G_u|$ vertices (Chemical Rubber Company, 1996, p. 201).
6. S_i is the (multi)set of vertex deleted subgraphs of G_i, and $|S_i| = |G_i| = |G_u|$.
7. A k-permutation of S_i is an ordered sequence of k (where $1 \leq k \leq |S_i|$) of the $|S_i|$ vertex-deleted subgraphs of G_i. The (multi)set of all possible k-permutations of S_i is denoted $k\text{-p}_{S_i}$. The number of k-permutations of S_i, i.e.

$|k\text{-p}_{S_i}|$, equals $\frac{|S_i|!}{(|S_i|-k)!}$ (Cormen, Leiserson, Rivest, & Stein, 2001, pp. 1095-1096). Traditionally, a k-permutation consists of a sequence of elements not occurring more than once in the sequence. In graph reconstruction, the elements, i.e. vertex-deleted subgraphs, would be unique if labeled; therefore, an ordering of vertex-deleted subgraphs constitutes a k-permutation. Nevertheless, the problem of reconstructing unlabeled vertex-deleted subgraphs may result in a k-permutation containing elements that appear to occur more than once in the sequence. Consequently, $k\text{-p}_{S_i}$ may also contain ‘repeat’ elements; therefore, $k\text{-p}_{S_i}$ may be considered a multiset.

8. The (multi)set of k-permutations of S_u is denoted $k\text{-p}_{S_u}$, and corresponds to the possible observation sequences of k of the $|S_u|$ vertex-deleted subgraphs of G_u.

9. For reasons previously discussed, some k-permutations of S_u may be isomorphic, and hence indistinguishable from the observer’s perspective since G_u is unknown. In this framework, observation sequence jk is the jth non-isomorphic k-permutation of S_u, and is denoted obseq_{jk} (Harary, 1964; Lauri, 2004). Furthermore, $m_S(e)$ denotes the multiplicity of element e in multiset S (Blizard, 1989; Bogart, 1983, p. 44; Hickman, 1980; Syropoulus, 2000; Wikipedia, 2008). Note that k represents the number of observations, i.e. the time step.

10. Implementing the notation above yields the following definition of accuracy:

$$\text{Prob}\{G_u = G_i |\ \text{obseq}_{jk}\} = \frac{m_{k\text{-p}_{S_i}}(\text{obseq}_{jk})}{\sum_i m_{k\text{-p}_{S_i}}(\text{obseq}_{jk})}$$

This definition of accuracy takes into account the fact that it may be possible to reconstruct more than one graph from the given observation sequence (Baldwin, 2003; Devore, 1987; Stockmeyer, 1976 as cited in Bondy, 1991). Taken together these graphs form the reconstruction family of the observation sequence (Klir, 1985). Determining the reconstruction family of an observation sequence has similarities with the legitimate subdeck and subdeck checking problems.

An example and associated results of this framework are provided for all 4 isomorphism classes for a simple graph with three vertices shown in Figure 4 (Chemical Rubber Company, 1996, p. 201). If $G_u = G_1$, then $j = 1 \ \forall \ k$. Table 31 displays the associated observation sequences and accuracies. In this case, the ally RN = adversary RN = duration = 3.

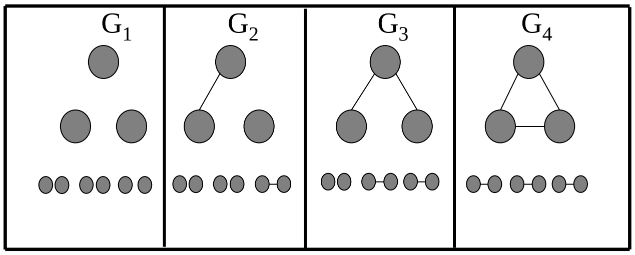
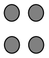
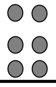


Figure 4 The Four Isomorphism Classes of a Three Vertex Graph

Table 31 Accuracy Statistics for $G_u = G_1$

obseq$_{1k}$		P(G$_u$=G$_1$ | obseq$_{1k}$)	P(G$_u$=G$_2$ | obseq$_{1k}$)	P(G$_u$=G$_3$ | obseq$_{1k}$)	P(G$_u$=G$_4$ | obseq$_{1k}$)
k = 1		0.5	0.33	0.17	0.0
k = 2		0.75	0.25	0.0	0.0
k = 3		1.0	0.0	0.0	0.0

If $G_u = G_2$, then there are $j = 2$ non-isomorphic observation sequences for $k = 1$, and $j = 3$ non-isomorphic observation sequences for $k = 2, 3$. Tables 32, 33 and 34 contain the observation sequences and associated accuracies for obseq$_{j1}$, obseq$_{j2}$ and obseq$_{j3}$, respectively.

The subgraph observation order is relevant to accuracy, which is in keeping with the notion of ally and adversary ordering (Bondy, 1991; Myrvold, 1988, 1990). Additionally, the ally RN = adversary RN = duration = 3 for $G_u = G_2$. This example

Table 32 Accuracy Statistics for $G_u = G_2$ and obseq$_{j1}$

obseq$_{j1}$	$P(G_u{=}G_1 \mid \text{obseq}_{j1})$	$P(G_u{=}G_2 \mid \text{obseq}_{j1})$	$P(G_u{=}G_3 \mid \text{obseq}_{j1})$	$P(G_u{=}G_4 \mid \text{obseq}_{j1})$
j = 1	0.5	0.33	0.17	0
j = 2	0.0	0.17	0.33	0.5

Table 33 Accuracy Statistics for $G_u = G_2$ and obseq$_{j2}$

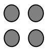
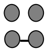
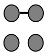
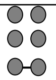
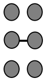
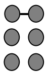
obseq$_{j2}$	$P(G_u{=}G_1 \mid \text{obseq}_{j2})$	$P(G_u{=}G_2 \mid \text{obseq}_{j2})$	$P(G_u{=}G_3 \mid \text{obseq}_{j2})$	$P(G_u{=}G_4 \mid \text{obseq}_{j2})$
j = 1 	0.75	0.25	0	0
j = 2 	0.0	0.5	0.5	0
j = 3 	0.0	0.5	0.5	0

Table 34 Accuracy Statistics for $G_u = G_2$ and obseq$_{j3}$

obseq$_{j3}$	$P(G_u{=}G_1 \mid \text{obseq}_{j3})$	$P(G_u{=}G_2 \mid \text{obseq}_{j3})$	$P(G_u{=}G_3 \mid \text{obseq}_{j3})$	$P(G_u{=}G_4 \mid \text{obseq}_{j3})$
j = 1 	0	1	0	0
j = 2 	0	1	0	0
j = 3 	0	1	0	0

illustrates that partial results can be misleading as evidenced by the accuracy values for G_1 and G_2 for obseq$_{1k}$ in the tables, i.e. G_1 is the most probable candidate for G_u until all subgraphs are revealed. Consequently, in this example, choosing the graph with the highest accuracy given partial information, i.e. $k < 3$ is not the best method for reducing the risk. An alternative method is to choose a member of the reconstruction family that minimizes error (Klir, 1985; G. J. Klir, personal communication, May 26, 2007). This follows from the fact that risk can be considered a measure involving both probability, i.e. accuracy, and severity (Haimes, 2004; Lowrance, 1976). Consider the case of obseq$_{11}$ with error defined as the symmetric difference between a graph and all other graphs in the reconstruction family (Banks & Carley, 1994; Kemeny, 1959). In this case, the reconstruction family consists of graphs G_1, G_2 and G_3, and the respective errors are 3, 2 and 3. Consequently, using only symmetric difference as a measure of risk, G_2 would be chosen as the least risk alternative for obseq$_{11}$. In the case of obseq$_{12}$, there is not a unique least error solution; hence one would possibly 'fall back' on accuracy as the selection method, which would lead to an incorrect conclusion regarding G_u. Nevertheless, studying both accuracy and error measurements may provide an analyst or decision-maker with valuable risk mitigation insight in cases where the graphs under consideration represent real world networks and systems.

If $G_u = G_3$, then there are $j = 2$ non-isomorphic observation sequences for $k = 1$, and $j = 3$ non-isomorphic observation sequences for $k = 2, 3$. The observation sequences and associated accuracies for obseq$_{j1}$ are the same as those in Table 32. Tables 35 and 36 contain the observation sequences and associated accuracies for obseq$_{j2}$ and obseq$_{j3}$, respectively. The subgraph observation order is again relevant to accuracy; furthermore, partial results can be misleading. The ally RN = adversary RN = duration = 3 for $G_u = G_3$.

Table 35 Accuracy Statistics for $G_u = G_3$ and obseq$_{j2}$

obseq$_{j2}$		P(G_u=G_1 | obseq$_{j2}$)	P(G_u=G_2 | obseq$_{j2}$)	P(G_u=G_3 | obseq$_{j2}$)	P(G_u=G_4 | obseq$_{j2}$)
j = 1	●–● ●–●	0.0	0.0	0.25	0.75
j = 2	● ● ●–●	0.0	0.5	0.5	0.0
j = 3	●–● ● ●	0.0	0.5	0.5	0.0

Table 36 Accuracy Statistics for $G_u = G_3$ and obseq$_{j3}$

obseq$_{j3}$		P(G_u=G_1 | obseq$_{j3}$)	P(G_u=G_2 | obseq$_{j3}$)	P(G_u=G_3 | obseq$_{j3}$)	P(G_u=G_4 | obseq$_{j3}$)
j = 1	●–● ●–● ● ●	0	0	1	0
j = 2	● ● ●–● ●–●	0	0	1	0
j = 3	●–● ● ● ●–●	0	0	1	0

If $G_u = G_4$, then $j = 1 \ \forall \ k$. Table 37 displays the associated observation sequences and accuracies. In this case, the ally RN = adversary RN = duration = 3.

Table 37 Accuracy Statistics for $G_u = G_4$

Obseq$_{1k}$		P(G_u=G_1 | obseq$_{1k}$)	P(G_u=G_2 | obseq$_{1k}$)	P(G_u=G_3 | obseq$_{1k}$)	P(G_u=G_4 | obseq$_{1k}$)
k = 1	●–●	0.0	0.17	0.33	0.5
k = 2	●–● ●–●	0.0	0.0	0.25	0.75
k = 3	●–● ●–● ●–●	0.0	0.0	0.0	1.0

This section provided a framework, with notation and a definition of accuracy, and results for the traditional reconstruction framework. From the results, it is seen that a large overlap between two graphs' vertex-deleted subgraphs contributes to the reconstruction accuracy. The overlap depends upon the diversity, i.e. number, of isomorphism classes of the vertex-deleted subgraphs generated by each graph. This concept illustrates how diversity can impact the number of possible orderings which in turn affects the duration, and the monotonicity of accuracy (Baldwin, 2003; Hemaspaandra *et al.*, 2007; McMullen, 2005; Myrvold, 1988; Stockmeyer, 1976 as cited in Bondy, 1991). The next section examines a framework for graph reconstruction when repeat observations of elements in S_u are permitted.

5.2.2 Framework for Modified Graph Reconstruction

The following constraints, definitions and notation comprise the framework applied to a modification of graph reconstruction; specifically, a framework in which subgraphs may be observed once or more than once; additionally, a subgraph or subgraphs may not be observed at all.

1. G_u: An undirected, reconstructable graph from which vertex deleted subgraphs are observed. G_u is unknown to the observer, and represents an unknown network structure.

2. $|G_u|$ is the number of vertices of G_u; also referred to as the order of G_u (Myrvold, 1990). $|G_u|$ is known to the observer.

3. S_u is the set of vertex deleted subgraphs of G_u, and $|S_u| = |G_u|$. Since the vertex deleted subgraphs in S_u may not be unique, S_u can be considered a multiset.

4. Each vertex deleted subgraph in S_u is denoted, in a manner similar to Harary (1964), as $V_{u,l}$ where $l = 1, \ldots, |S_u|$. Furthermore, each $V_{u,l}$ may be observed an *a priori* unknown number of times, $r_{V_{u,l}}$.

5. G_i is the ith isomorphism class of a graph with $|G_u|$ vertices.

6. S_i is the set of vertex deleted subgraphs of G_i, and $|S_i| = |G_i| = |G_u|$.

7. A string of S_i is a sequence of vertex-deleted subgraphs in S_i (by definition, elements in a string may be repeated). A string of length k is referred to as a

k-string. The (multi)set of all possible k-strings of S_i is denoted $k\text{-st}_{S_i}$. The number of k-strings of S_i, i.e. $|k\text{-st}_{S_i}|$, equals $|S_i|^k$ (Cormen *et al.*, 2001, p. 1095).

8. The set of k-strings of S_u is denoted $k\text{-st}_{S_u}$, and corresponds to the possible observation sequences containing k observations. Since repeat observations are permissible, it is possible that $k \geq |S_u|$. Each observation in a k-string of S_u corresponds to a vertex-deleted subgraph, i.e. $V_{u,l}$; furthermore, $r_{V_{u,l}} \leq k \; \forall \; l, k$ (Hemaspaandra *et al.*, 2007, p. 114).
9. Some k-strings of S_u may be isomorphic, and hence indistinguishable from the observer's perspective since G_u is unknown. Observation sequence jk is the jth non-isomorphic k-string of S_u, and is denoted obseq_{jk}.

This framework can be considered a generalization of the traditional (and set) reconstruction framework/problem. If vertex deleted subgraphs of G_u are being observed and $k \leq |G_u|$ and $r_{V_{u,l}} \leq 1 \; \forall \; l$, then the modified framework is identical to the traditional reconstruction framework/problem. If non-isomorphic vertex deleted subgraphs of G_u are being observed and k is less than or equal to the number of non-isomorphic vertex deleted subgraphs of G_u and $r_{V_{u,l}} = 1 \; \forall \; l$ s.t. l is the numerical identifier associated with a non-isomorphic vertex deleted subgraph, then the modified framework can represent the set reconstruction problem.

Other graph reconstruction aspects can also be addressed by the extended framework. For example, since k represents the total number of observations, which impacts the number of possible observations of each vertex-deleted subgraph, (ally) reconstruction numbers can be addressed as follows: Determine the minimum number of observations, k, of vertex deleted subgraphs, i.e. $V_{u,l}$, that must be observed in order to reconstruct G_u given that $r_{V_{u,l}} \leq 1$; consequently, $k \leq |G_u|$, assuming the reconstruction conjecture holds (Harary & Plantholt, 1985; Hemaspaandra *et al.*, 2007, p. 114).

Implementing the notation above yields the following definition of accuracy (conceptually similar to a function describing the behavior of a generative system as provided in Klir (1985)) for the modified reconstruction framework in which all, not just the non-isomorphic, vertex-deleted subgraphs may be observed:

$$\text{Prob}\{G_u = G_i | \text{ obseq}_{jk}\} = \frac{m_{k\text{-st}_{S_i}}(\text{obseq}_{jk})}{\sum_i m_{k\text{-st}_{S_i}}(\text{obseq}_{jk})}$$

An example and associated results of the modified framework are provided for all 4 isomorphism classes for a simple graph with three vertices shown in Figure 4. If $G_u = G_1$, then $j = 1 \; \forall \; k$; however, since repeat observations are permitted $0 \leq r_{V_{u,l}}$. Additionally, k could be greater than 3 since repeat observations are permitted; however, for illustrative purposes, Table 38 displays the associated observation sequences and accuracies for $1 \leq k \leq 3$.

Table 38 Modified Accuracy Statistics for $G_u = G_1$

Obseq_{1k}		$P(G_u{=}G_1 \mid \text{obseq}_{1k})$	$P(G_u{=}G_2 \mid \text{obseq}_{1k})$	$P(G_u{=}G_3 \mid \text{obseq}_{1k})$	$P(G_u{=}G_4 \mid \text{obseq}_{1k})$
k = 1	● ●	0.5	0.33	0.17	0.0
k = 2	● ● ● ●	0.64	0.29	0.07	0.0
k = 3	● ● ● ● ● ●	0.75	0.22	0.03	0.0

For all k, G_u cannot be unambigously reconstructed as G_1; therefore, the duration is considered infinite (Hemaspaandra *et al.*, 2007). Nevertheless, it is most probable, given the observations, that $G_u = G_1$. As previously discussed, error could also be evaluated in order to form a conclusion that addresses risk.

If either $G_u = G_2$ or $G_u = G_3$, then there are $j = 2$ non-isomorphic observation sequences for $k = 1$, $j = 4$ non-isomorphic observation sequences for $k = 2$ and $j = 8$ non-isomorphic observation sequences for $k = 3$. Tables 39, 40 and 41 contain the observation sequences and associated accuracies for obseq$_{j1}$, obseq$_{j2}$ and obseq$_{j3}$, respectively.

Table 39 Modified Accuracy Statistics for $G_u = G_2$ and obseq$_{j1}$

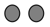
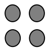
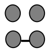
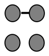
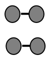
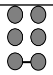
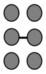
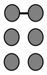
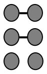
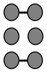
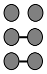
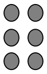
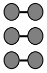
obseq$_{j1}$	$P(G_u=G_1 \mid \text{obseq}_{j1})$	$P(G_u=G_2 \mid \text{obseq}_{j1})$	$P(G_u=G_3 \mid \text{obseq}_{j1})$	$P(G_u=G_4 \mid \text{obseq}_{j1})$
j = 1 	0.5	0.33	0.17	0
j = 2	0.0	0.17	0.33	0.5

Table 40 Modified Accuracy Statistics for $G_u = G_2$ and obseq$_{j2}$

obseq$_{j2}$	$P(G_u=G_1 \mid \text{obseq}_{j2})$	$P(G_u=G_2 \mid \text{obseq}_{j2})$	$P(G_u=G_3 \mid \text{obseq}_{j2})$	$P(G_u=G_4 \mid \text{obseq}_{j2})$
j = 1 	0.64	0.29	0.07	0.0
j = 2 	0.0	0.5	0.5	0.0
j = 3 	0.0	0.5	0.5	0.0
j = 4 	0.0	0.07	0.29	0.64

The subgraph observation order is relevant to accuracy; however, since repeat observations are permitted, accuracy values can be misleading. Additionally, the above example illustrates, in the context of repeated observations, why the set reconstruction conjecture holds only for graphs with four or more vertices. Regardless of the number of observations, if G_u is either G_2 or G_3, one cannot distinguish between the two graphs (Bondy, 1978; Harary, 1964). Consequently, the duration is infinite when G_u is either G_2 or G_3. In similar applications where a system cannot be reconstructed given current observations, it may be reasonable to choose a threshold accuracy value that, if exceeded, will be considered sufficient to claim $G_u = G_i$, i.e. continue observations until accuracy is greater than or equal to some value. For example, assume a threshold value of 0.8 was chosen; furthermore, assume the current observations are those in obseq$_{73}$ of Table 41. If the next observation is composed

Table 41 Modified Accuracy Statistics for $G_u = G_2$ and obseq$_{j3}$

obseq$_{j3}$	$P(G_u{=}G_1 \mid \text{obseq}_{j3})$	$P(G_u{=}G_2 \mid \text{obseq}_{j3})$	$P(G_u{=}G_3 \mid \text{obseq}_{j3})$	$P(G_u{=}G_4 \mid \text{obseq}_{j3})$
j = 1 	0.0	0.67	0.33	0.0
j = 2 	0.0	0.67	0.33	0.0
j = 3 	0.0	0.67	0.33	0.0
j=4 	0.0	0.33	0.67	0.0
j=5 	0.0	0.33	0.67	0.0
j=6 	0.0	0.33	0.67	0.0
j=7 	0.75	0.22	0.03	0.0
j=8 	0.0	0.03	0.22	0.75

of another set of two vertices that are not connected, then the accuracies associated with G_1, G_2 and G_3 would be 0.83, 0.16 and 0.01, respectively. Consequently, G_1 would be chosen as the reconstruction of G_u based on exceeding the threshold value of 0.8.

If $G_u = G_4$, then $j = 1 \ \forall \ k$. Table 42 displays the associated observation sequences and accuracies for $1 \leq k \leq 3$. For all k, G_u cannot be unambigously reconstructed as G_4; therefore, the duration is considered infinite. Nevertheless, it is most probable, given the observations, that $G_u = G_4$.

The following example considers the four vertex graph, G_1, in Figure 5, to be G_u, and provides a finite duration reconstruction within the modified framework. There are eleven classes of graphs (containing four vertices) which are not isomorphic (West, 2001, p. 11). Figure 5 contains graphs from four of the classes based on possible reconstructions given the observation sequence shown in Table 43. A few remarks are in order. First, accuracy may be misleading given partial information;

Table 42 Modified Accuracy Statistics for $G_u = G_4$

Obseq_{1k}		$P(G_u=G_1 \mid \text{obseq}_{1k})$	$P(G_u=G_2 \mid \text{obseq}_{1k})$	$P(G_u=G_3 \mid \text{obseq}_{1k})$	$P(G_u=G_4 \mid \text{obseq}_{1k})$
k = 1	●–●	0.0	0.17	0.33	0.5
k = 2	●–● ●–●	0.0	0.07	0.29	0.64
k = 3	●–● ●–● ●–●	0.0	0.03	0.22	0.75

however, partial information can aid in constraining the reconstruction family. For example, the first observation in Table 43 reduces the number of possible reconstructions from all eleven classes to four classes. Consequently, partial information can provide a measure of accuracy. Second, the observation sequence impacts accuracy (magnitude and monotonicity) and duration ('short', 'long', infinite). These two concepts are interrelated and in accordance with the notion of reconstruction information/numbers (Harary & Plantholt, 1985; Manvel, 1969; Myrvold, 1988). Lastly, if the observation sequence had been slightly modified, i.e. three isolates, followed by the path subgraph, followed by the connected dyad with the isolate, the result would have been the same as a set reconstruction attempt.

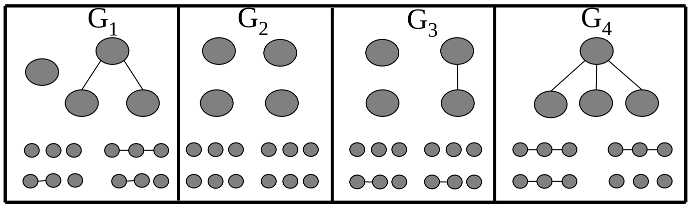
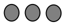
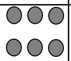
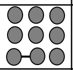
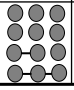


Figure 5 Four Graphs with Four Vertices

Given the modified framework, where repeat observations of all vertex deleted subgraphs are possible (and consequently, the possibility exists that not every subgraph will be observed), it is conjectured that the only cases in which reconstruction of G_u, with $|G_u| > 3$, can be guaranteed are those in which all appropriate non-isomorphic vertex-deleted subgraphs have been observed. In order for the observed non-isomorphic vertex-deleted subgraphs to be 'appropriate', the non-isomorphic el-

Table 43 Accuracy Statistics for Finite Duration Reconstruction

$obseq_{1k}$	$P(G_u=G_1 \mid obseq_{jk})$	$P(G_u=G_2 \mid obseq_{jk})$	$P(G_u=G_3 \mid obseq_{jk})$	$P(G_u=G_4 \mid obseq_{jk})$
k = 1 	0.125	0.5	0.25	0.125
k = 2 	0.045	0.73	0.18	0.045
k=3 	0.33	0.0	0.67	0.0
k=4 	1.0	0.0	0.0	0.0

ements of S_u can not be contained in any other single S_i. This constraint is necessary because the possibility of repeat observations would preclude one from determining whether G_u was the true graph, or if G_i was the true graph but all non-isomorphic vertex deleted subgraphs of G_i had not yet been observed. This constraint is related to, but more stringent than, that imposed by the set reconstruction conjecture. Given the conjecture for the modified framework, the only graph in Figure 5 that can be reconstructed is G_1; furthermore, reconstruction is only possible if all non-isomorphic elements, i.e. vertex deleted subgraphs, of G_1 have been observed. Consequently, both constraints (i.e. observation of all non-isomorphic vertex deleted subgraphs in S_u, and the non-isomorphic subgraphs not being a subset of any other S_i) form a necessary and sufficient condition for reconstruction in the modified framework.

This section provided a framework, with notation and a definition of accuracy, and results for the modified reconstruction framework in which repeat observations of the vertex deleted subgraphs are permitted. Examples illustrating the modified framework and associated accuracy and duration concepts were presented. As in traditional reconstruction, the observation sequence impacts the accuracy and duration.

5.3 Conclusion

This chapter provided an extension to network reconstruction that encompasses possible repeat observations of vertex deleted subgraphs, as well as possible non-observations. Traditional reconstruction poses reconstruction in terms of the number of subgraphs required to reconstruct the original network. This research associates that number, in a time context, i.e. the duration to reconstruct, since repeat observations are permissible. Additionally, stochasticity is addressed via the definition of accuracy introduced in the modified framework. Both accuracy and duration are affected by the sequence of observations, as well as, the set of non-isomorphic vertex deleted subgraphs. The modified framework is a more general formulation for network reconstruction, and can address an, arguably, more difficult reconstruction problem. In the next chapter, the reconstruction of causal networks is addressed. Additionally, a method for determining plausible social influence network structures, and hidden individuals therein, is presented.

6. Social Influence Networks

6.1 Introduction

This chapter introduces a method that couples social network influence concepts with causal exploratory analysis to facilitate determining social influence network (SIN) structures and revealing hidden individuals within such networks. The method is demonstrated using real-world data. Additionally, incorporating causality concepts into network reconstruction is addressed.

6.2 Background

Social influence networks are networks composed of individuals who may exert influence on each other. Causal exploratory analysis is the process of obtaining causal relations among variables of interest from empirical data. Based on concepts discussed in the literature review, e.g. asymmetry of influence, causal networks are chosen to represent influence networks, and the causal networks are represented as directed acyclic graphs (DAGs) (March, 1955, pp. 436-437). Furthermore, in certain cases, it may be appropriate to assume the network nodes act in concert for some purpose, so isolates are not present; hence, the DAGs would be connected. While the focus in this section is on social influence networks, the methods presented can be applied to a variety of other networks that operate on the principle of a node causing effects on or influencing another node or nodes. The contributions associated with this portion of the research are:

1. A method, based on causal exploratory analysis, to determine candidate structures of social influence networks.
2. Detection of unrevealed individuals in social influence networks using causality principles.

3. A method that formalizes the integration of graph reconstruction with causal exploratory analysis and provides associated accuracy metrics.

6.3 Methodology

The following sections provide the methods for each contribution.

6.3.1 Social Influence Network Exploratory Analysis

In order to explore various social influence network structures, it is important to identify both influencing and influenced individuals. Often these individuals (and possibly some of the influence relationships among them) are known prior to analysis. Just as in Bayesian belief networks (Cooper & Herskovits, 1992), prior relational knowledge can be incorporated in social network analysis (Doreian, 2001; Friedkin, 1998). The prior relational knowledge has been derived from, for example, ties between individuals and subsequently embedded in and expressed via the **W** matrix of the network model, $\mathbf{y} = \alpha\mathbf{W}\mathbf{y} + \beta\mathbf{X}\mathbf{b} + \mathbf{u}$, where **W** represents the SIN (Friedkin, 1990, 1998). In this research, exploratory analysis is used in an attempt to identify, from empirical data and prior (but usually incomplete) knowledge, influence relationships among known individuals. In fact, the resulting causal relationship structure represents the SIN and could be loosely considered a proxy for **W** by modifying certain constraints given in Friedkin (1986). Furthermore, exploratory analysis can lead to the revelation of previously unconsidered or unknown individuals. Another requirement of exploratory analysis, in the current context, is identifying a measurement or measurements of causality/influence so that probabilistic independencies between the nodes can be calculated (Geiger & Pearl, 1990, pp. 3, 10; Pearl 1988, pp. 81-86, 89 91-94, 116-119, 122; Shipley, 2002, pp. 8-9, 36-37, 90-94; Spirtes *et al.*, 2000, pp. 43-44, 82, 139; Verma & Pearl, 1990, p. 71; Verma & Pearl, 1991, pp. 256, 264). Independency calculations between nodes of causal models representing some network types (e.g. engineered networks) are, generally, not too difficult; however, in social networks, more scrutiny is required. As noted by Spirtes *et al.* (2000), the causal Markov condition may not hold when proxies are used in causal analysis (cf. Salmon (1984)); nevertheless, this is foundational research to which other actions, such as appropriately defining variables may eliminate such errors (Spirtes *et al.*, 2000, p. 37). As previously discussed, this research will examine a single event category, i.e. a single relation, in exploratory analysis.

The following approach is demonstrated for exploratory analysis of a SIN using a single relation.

1. Identification of a set of individuals.
2. Selection of an event category used to measure influence between individuals in the social network.
3. Production of candidate SIN structures from causal exploratory analysis of empirical data. In this dissertation, causal modeling software, Tetrad IV, is employed (Glymour, Scheines, Spirtes, & Ramsey, 2004b).

As discussed in the literature review, the causal inference algorithms may output a partially oriented graph; consequently, the true causal structure is not uniquely determined (Richardson, 1996; Shipley, 2002, pp. 256-260; Silva, 2005, pp. 9, 24; Spirtes *et al.*, 1993, pp. 180-183; Spirtes *et al.*, 2000, pp. 6, 59, 61, 82-87, 139-140; Zhang, 2004). Nevertheless, one could derive graphs from the output, and, if desired, compare them against a conjectured 'ground truth' SIN structure (Glymour *et al.*, 2004b; Joreskog & Sorbom, 1995, p. 22; Shipley, 2002, pp. 102-103; Spirtes *et al.*, 2000; Verma & Pearl, 1991). In order to measure the difference between the proposed 'ground truth' graph, g_1, and a graph, g_2, derived from the output, a function based on a metric given in Banks and Carley (1994) is employed. The metric provided in Banks and Carley (1994) is: $d^+(g_1, g_2) = tr[(\mathbf{G}_1 - \mathbf{G}_2)^T(\mathbf{G}_1 - \mathbf{G}_2)]$ where $\mathbf{G}_i$ is the adjacency matrix of g_i. According to Banks and Carley (1994), this is a directed network variant of the Kemeny (1959) metric. In order to address latent variables, the method in this research assumes that each double-headed arrow adds one unit of distance. For example, if the output SIN contains a single double-headed arrow, then the distance from the 'ground truth' network is the directed network metric result for the network variables which do not contain the double-headed arrow, plus one unit. Figure 6 provides an illustrative example for two different cases.

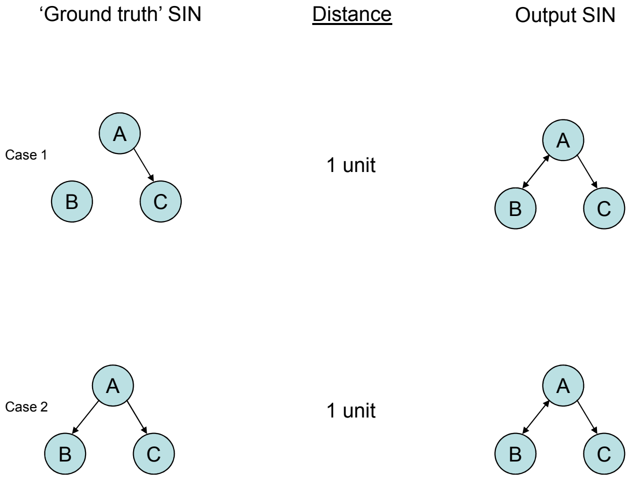


Figure 6 Distance Function Example

6.3.2 Detecting Unrevealed Individuals in Social Influence Networks

Results from exploratory analysis may contain a bi-directed arc between two variables. One possible interpretation of such an arc is the presence of a hidden variable that directly causes the original two variables (Glymour *et al.*, 2004a, 2004b; Shipley, 2002, pp. 26, 256, 266-267; Spirtes *et al.*, 2000, pp. 125, 144-145; Verma & Pearl, 1991). In the single relation SIN previously discussed, such a variable indicates a hidden individual, in that event category, directly influencing two known individuals in the SIN. This is due to the assumption (based on the reasoning provided by March) that the association between events corresponds to influence between individuals in that particular activity. (March, 1953-54, pp. 469-470; March, 1955, pp. 435-436) Consequently, revealing a hidden individual in a SIN is accomplished by simply performing exploratory analysis, and examining the results for appropriate indicators. In order to validate this approach, the event category data for an individual in the SIN is removed, subsequent exploratory analysis is conducted, and the output is examined for appropriate hidden variable indicators (Shipley, 2002, pp. 266, 267; Spirtes *et al.*, 2000, pp. 144-145; Glymour *et al.*, 2004a, 2004b).

6.3.3 Causal Graph Reconstruction

This section addresses the coupling of graph reconstruction with causality concepts. Consider a DAG, DAG_u, that represents a causal network. Since causal network nodes influence other nodes, DAG_u contains labeled nodes so that meaningful causal analysis can be performed (Heise, 1975, pp. 39-40, 45; Verma & Pearl, 1991, pp. 256). Harary and Manvel (1970) proved that every graph with at most two unlabeled vertices require no more than three of its vertex deleted subgraphs in order to be reconstructed. Consequently, reconstruction of a fully labeled causal graph, i.e. a social influence network, is possible; however, ambiguity regarding the structure of DAG_u may arise if less than three subgraphs are provided. Furthermore, the notion of equivalent causal structures complicates the situation (Verma & Pearl, 1991, pp. 256).

The following approach will be employed to develop a framework for reconstructing labeled DAGs representing causally sufficient networks, i.e. networks with no latent common causes of variables in the variable set of interest (Shipley, 2000, pp. 259-260, 264-266; Spirtes *et al.*, 2000, p. 22). The associated method assumes the existence of a stable, stationary causal structure having reached a static value at the observation time, and an observer is given parts of the structure, i.e. the observer is not required to perform exploratory analysis to determine the causal substructure. (Heise, 1975, pp. 48-49, James *et al.*, 1982, p. 49) The causal structure is unknown to the observer, but the number, $n \geq 3$, of nodes is known. Additionally, each causal subgraph is given once and only once.

1. At each time step, $t_i, i = 1, 2, 3$, a vertex deleted subgraph dag_{t_i} containing $n-1$ labeled nodes and their associated edges, is provided. Additionally, there are $n!$ possible sequences (denoted $provseq_j, j = 1, 2, \ldots, n!$) in which subgraphs can be provided; however, per Harary and Manvel (1970), the first three (or less) subgraphs of each sequence suffice to reconstruct DAG_u. Consequently, at a maximum, there are $n(n-1)(n-2)$ relevant sequences.

2. Determine the set of possible reconstructions, $poss\_rec_{t_i}$ given the subgraphs viewed through t_i; note that $poss\_rec_{t_3} = DAG_u$. A related measure of accuracy, i.e. the probability of choosing the correct causal graph given the subgraphs viewed through t_i, is defined as $part\_acc_{t_i} \equiv |poss\_rec_{t_i}|^{-1}$. This definition of accuracy is derived from the notions of reconstruction family and uniform prior probabilities (Cooper & Herskovits, 1992; Klir, 1985).

3. After any time step, causal exploratory analysis (i.e. independence tests) may be performed for the n nodes to determine the set of equivalent causal graphs, $pats$. The set $(poss\_rec_{t_i} \cap pats)$ contains equivalent causal graphs consistent with the causal substructures provided through t_i. Consequently, an associated measure of accuracy, i.e. the probability of choosing the correct causal graph given the subgraphs viewed through t_i and exploratory analysis of the n nodes, can be derived as follows: $aug\_acc_{t_i} \equiv |(poss\_rec_{t_i} \cap pats)|^{-1}$.

4. Continue this process until the graph has been reconstructed or a desired number of time steps has been completed.

The result of this approach is a measure or understanding of the feasible solutions, the variety of such solutions and, hence, the amount of uncertainty regarding possible solutions, i.e. social influence networks (Michalewicz & Fogle, 2002). Additionally, insight is gained regarding the value of prior knowledge, i.e. the provided causal substructures, and the timing of independence tests. The value of prior knowledge may help identify where one might try to intervene in the network, in order to obtain further information on the true causal structure (Shipley, 2002, pp. 258-

260). These considerations can aid analysts and decision makers in assessing risks associated with accepting a particular reconstruction as the solution.

An alternative but related approach/perspective is to perform independence tests before any causal substructures are provided. Subsequently varying the sequence of provided causal substructures should give insight (as did the previous approach) regarding the contributory value of each causal substructure in accurately determining the true causal graph. In this approach, the definition of $poss\_rec_{t_i}$ is slightly modified, and becomes the set of possible reconstructions given both the subgraphs viewed through t_i and the result of the independence tests. The modified $poss\_rec_{t_i}$ produces a modified $part\_acc_{t_i}$ that is equivalent to $aug\_acc_{t_i}$; hence, one could also derive the information from this alternative perspective directly from $aug\_acc_{t_i}$.

6.4 Data and Results

Based on ideas from Bullock and Brady (1983) and March (1955), the SIN analysis data consisted of five United States Senators: Senator Reid, Senator Biden, Senator Harkin, Senator Brown and Senator Cardin. The event category chosen was the voting activity of Senate members. The roll call votes of the five Senators during the 110th Congress from 1 January 2007 - 19 Jun 2008 comprised the data. There were 596 such votes; however, not all five Senators always voted (United States Senate, 2007, 2008a). In such cases, these data points were removed resulting in a final data set of 383 votes. The rationale for choosing these five Senators included the fact that their positions indicated a somewhat apparent chain of command, i.e. a potential 'ground truth' network. Additionally, the choice was made in order to permit causal exploratory analysis; consequently, the individuals were chosen on a notion that they related to one another asymmetrically with respect to influence; therefore, the SIN could be represented by a DAG and examined via causal analysis. Future research could examine relaxing the asymmetrical influence assumption (cf. Friedkin

(1986)) via algorithms for causal exploratory analysis that address directed cycles (Richardson, 1996). Senator Reid was the Senate majority leader, while Senators Brown and Cardin were freshmen (United States Senate, 2008b, 2008c). Senators Biden and Harkin were chairmen of committees of which Senators Cardin and Brown, respectively, were members. Additionally, Senators Biden and Harkin were members of one other committee to which Senators Cardin and Brown, respectively, were members (Senate of the United States, 2007a, 2007b). Another factor influencing the choice of these five Senators was the desire to control for political party since it was assumed that influence within the same party would be greater than across parties. Consequently, the five Senators chosen belong to the Democratic party (United States Senate, 2008d). The network 'ground truth' was characterized as a hierarchical chain of command as shown in Figure 7. It should be noted, however, that this illustrative example is a simplification of the complexities of the U.S. Senate. In studying the detection of hidden individuals in a SIN, the data on the Senators was used. The data set to illustrate causal graph reconstruction consists of connected, labeled DAGs with three vertices and two edges; however, the analysis approach can be applied to larger, labeled DAGs.

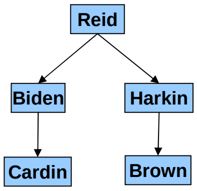


Figure 7 Chain of Command for Five Senators

6.4.1 Assumptions and Limitations

In order to apply the FCI algorithm, it is necessary to define random variables, i.e. nodes. For the illustrative data analyzed in this dissertation, a node represents a Senator's votes, the units are individual roll call votes and the possible values are yea and nay. These representations were chosen for clarity and causal contextual reasons as given in Cooper (1999), Howard (1988), Scheines, Spirtes, Glymour, Meek, & Richardson (1995) and Spirtes *et al.* (2000). The FCI algorithm, used in this dissertation to accomplish causal exploratory analysis, requires the Causal Markov and faithfulness conditions to hold (for the distribution to which the measured variables belong); otherwise incorrect results are possible (Cooper, 1999; Scheines *et al.*, 1995; Spirtes *et al.*, 2000). Spirtes *et al.* (2000) provide a discussion of the conditions, and when it is inappropriate to assume they hold (elaborations of the concepts are given in Scheines *et al.* (1995) and Shipley (2000)) (Spirtes *et al.*, 2000, pp. 32-42, 124, 296). Spirtes *et al.* (2000) state that,

> The Causal Markov Condition does not apply to systems of variables in which some variables are defined in terms of other variables, nor to systems with interunit causation ...even when [the Causal Markov Condition] is true of the population described by some data-generating process, it may not characterize the conditional independence relations found for measured variables in a sample due to:
>
> 1. sampling error;
> 2. causal relations between the sampling mechanism and the observed variables...;
> 3. lack of causal sufficiency among the measured variables...;
> 4. aggregation of variable values...;
> 5. when one variable is a function of another variable by definition...;
> 6. samples in which for some units A causes B and for other units B causes A;
> 7. reversible systems.

(p. 296)

Varying results are also possible if the algorithm is implemented with different alpha values (Shipley, 2002; Spirtes *et al.*, 2000, p. 351). The research objectives, i.e. either attempting to derive many potential structures, or high confidence in the resulting structures, will dictate the choice of the alpha value and affect the power of involved tests (Devore, 1987, pp. 277-287; Law & Kelton, 2000, pp. 257-258; Scheines *et al.*, 1995; Shipley, 2002; Spirtes *et al.*, 2000, pp. 115-121, 204-205).

This concept is examined in the following section. The Faithfulness condition can be violated in mixed distributions and when variables have relationships that are deterministic (Spirtes *et al.*, 2000, pp. 39-40, 53). For the method presented in this dissertation, it was assumed both the Causal Markov and Faithfulness conditions held. With respect to the illustrative example, the variables are not defined in terms of each other; however, there is some interunit causation (Spirtes *et al.*, 2000; Sober, 1988). There may also exist other issues corresponding to the above concepts and itemized list, as well as other potentially problematic issues (e.g. results when discrete variables are used as discussed in Devore (1987) Fienberg (1977, 2007), Larntz (1978) and MacDonald (2008)) (Spirtes *et al.*, 2000, pp. 95, 140, 351; Scheines *et al.*, 1995). Nevertheless, the research presented here is focused on introducing a method to assist in identifying SIN structures and reveal hidden individuals in a SIN. Future research could examine the viability of analyzing a SIN (and obtaining satisfactory data) in the manner presented in this dissertation.

Regarding the interpretation of FCI output and its application in this research, some concepts and associated interpretations follow. First, according to Glymour *et al.* (2004a), an arrow from one variable to another, "... indicates that there is a causal pathway ... connecting the two variables ... It does not necessarily mean that in the true causal graph, the connected variables have a direct causal connection." (Spirtes, personal communication, October 2008; Glymour *et al.*, 2004a, p. 113). This concept is reiterated in discussions of direct cause and specifying associated mechanisms (Scheines *et al.*, 1995; Shipley, 2000; Spirtes *et al.*, 2000). Alternatively, Merton (1957) defined interpersonal influence in terms of direct interaction between individuals (and one could specify this as forceful physical interaction, e.g. as given in the account of Seattle riots by Gillham and Marx (2000) and the definition in Merton (1949)); furthermore, Merton (1949) specifically did not concern himself with aspects of political and administrative power (on masses of individuals) in his study of interpersonal influence in a community. Notwithstanding the role of

Merton's interpersonal influence definition and research in forming concepts of this dissertation, this research does not limit interpersonal influence to direct interaction. Consequently, an arc from one individual to another represents the existence of an influence pathway, i.e. relationship, from the former to the latter, and this may occur through unobserved (and in some cases, observed) individuals (Richardson, 1996; Scheines *et al.*, 1995; Scheines, Spirtes, Glymour, Meek, & Richardson, 1998; Shipley, 2000; Spirtes *et al.*, 2000; Spirtes, personal communication, October 2008).

It is certainly possible that there are no unrevealed individuals (mediating or otherwise), in the social network under examination. If this is the case, then an arc can be interpreted akin to the interpersonal influence definition in Merton (1957) (Scheines *et al.*, 1995; Shipley, 2000; Spirtes *et al.*, 2000). It is also possible that there are mediating individuals (in a directed path from the former to the latter) who are unrevealed; such hidden individuals are not detectable given the method presented in this research. If it was desired to examine interpersonal influence in the sense of Merton (1957), one could still apply this method but a constrained and scrutinized data set would be required.

Second, as indicated in Merton (1949), some may not consider the illustrative example's political arrangement as a SIN. This research employs the Wasserman and Faust (1994) definition of a social network, which is rather broad; therefore, the political arrangement is considered a SIN. Regardless of one's definition of a social network, the contribution of the method in this dissertation stands.

Third, as discussed in previous chapters and assumed in this research, influence is examined vis-*á*-vis causal exploratory analysis. Causal analysis does not necessarily yield results one might consider in the realm of 'traditional' influence. Consider the illustrative example: causal exploratory analysis does not require a yea vote by one member to yield a yea vote by another member (as in the discussion on voting in March (1955)); a plausible interpretation could be that a yea vote by one member yields a nay vote by the other member (Spirtes *et al.*, 2000, p. 21).

Consequently, one might consider such results as counter-influence vice 'traditional' influence (as possibly interpreted from the definition of interpersonal influence in Merton (1957)). This dissertation assumes the more inclusive view of influence, i.e both counter-influence and 'traditional' influence are realizable.

6.4.2 Social Influence Network Exploratory Analysis

Causal exploratory analysis of the Senator data yielded the structure in Figure 8. The analysis was performed using Tetrad's FCI algorithm, which produces causal structures based on independence tests among the variables. The null hypothesis for such tests is that (sets of) variables are (conditionally) independent, and $\alpha = 0.05$ for the analysis resulting in Figure 8 (Pearl, 2001; Spirtes *et al.* 2000; Verma & Pearl, 1991). Based on simulation results mentioned in Spirtes *et al.* (2000), the independence test employed was the G^2 test vice the X^2 test (Spirtes *et al.*, 2000, p. 95). Shipley (2000) and Scheines *et al.* (1995) (in part based on tests by Shipley (1997) and those recorded in Spirtes *et al.* (2000)) recommend varying the value of α; consequently, additional analyses were conducted for α values of 0.1, 0.15, 0.2 and 0.3. The first three values produced the output shown in Figure 9, while the 0.3 level produced the graph in Figure 10. The differences in the structures of Figures 9 and 10 are edge orientations (or the lack thereof), and are likely due to the level at which the independence tests were conducted, which Spirtes *et al.* (2000) cited as a factor that could affect FCI output (Spirtes *et al.*, 2000, p. 351). As expected according to the null hypothesis, the number of edges increased with the significance level (Devore, 1987; Scheines *et al.*, 1995). In the output of FCI, an edge between variables indicates some (causal) association, and as indicated in Figure 8, edges may have circles at the endpoints, where a circle indicates that an arrowhead may or may not be present subject to certain constraints, e.g. no cycles (Glymour *et al.*, 2004a; Meek, 1995; Pearl, 1988; Shipley, 2002; Spirtes *et al.*, 2000). Consequently, at least two equivalent causal graphs, shown in Figure 11, can be derived from the $\alpha = 0.05$ output.

If background knowledge had been provided, the search of candidate structures and resulting output could be reduced (Spirtes *et al.*, 2000). The FCI algorithm permits the possibility that latent variables (representing common causes) may exist in the data, and in the illustrative example's output, this possibility is realized. In such cases, there is an unlimited number of equivalent structures due to infinite representations possible with latent variables (Scheines *et al.*, 1998; Silva, 2005). If the number of latent variables was constrained as in Zhang (2004), the number of equivalent causal structures would be finite (Silva, 2005, pp. 9, 24).

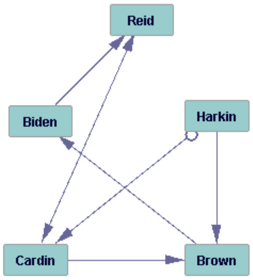


Figure 8 Causal Exploratory Results for Five Senators, $\alpha = 0.05$

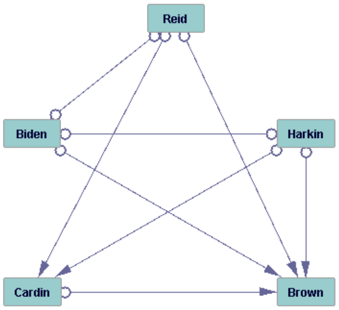


Figure 9 Causal Exploratory Results for Five Senators, $\alpha = 0.1 = 0.15 = 0.2$

Some interesting insights can be extracted from the output for $\alpha = 0.05$. First, there is no direct influence relationship between Senators Reid and Harkin. Second, there is an influence relationship from Senator Harkin (a committee chairman) to Senator Brown (a respective committee freshman), and a possible influence relation-

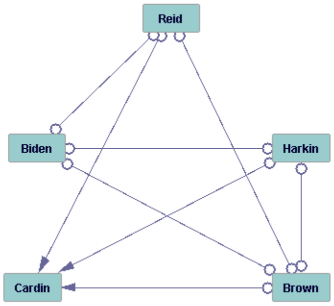


Figure 10 Causal Exploratory Results for Five Senators, $\alpha = 0.3$

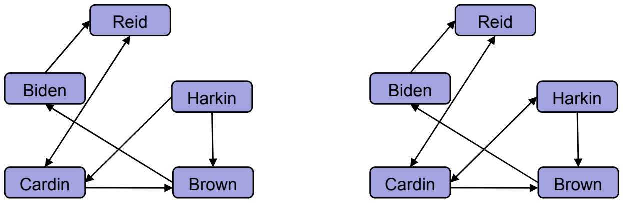


Figure 11 Causal Exploratory Derivatives for Five Senators, $\alpha = 0.05$

ship from Senator Harkin (a committee chairman) to Senator Cardin (a freshman). The analogous relationships are not present between Senators Biden, Cardin and Brown. An interesting note is that it was not originally realized that Senators Cardin and Harkin were co-members on the same committee; however, the analysis identified an influence relationship between the two Senators. Third, there exists an influence relationship from a freshman (Senator Brown) to a chairman (Senator Biden) of a different committee. Lastly, there does exist influence between freshmen, i.e. from Senator Cardin to Senator Brown. Not all insights are intuitive, and further investigation could be accomplished in such areas.

As evidenced in Figure 8, there is disparity between the proposed 'ground truth' and exploratory structures. The baseline disparity includes one edge oriented differently, two absent edges and four additional edges. From this baseline, further discrepancies can arise due to edge direction of partially oriented edges. Under the assumption that a double-headed arrow always adds one unit of distance, the minimum distance from the two potential structures to the 'ground truth' network is eight units. As a reference, the maximum possible distance from the ground truth network is fourteen units without any causal (orientation) constraints, given the above assumption (Glymour *et al.*, 2004a; Meek, 1995; Spirtes *et al.*, 2000). Multiple reasons could account for the disparity, such as the need to include other event categories or characteristics (e.g. personal attributes such as conservative or liberal; cf. Batt (2002), Project Vote Smart (2008), Smith (2001) and United States Senate (2008e)), an incorrect ground truth structure, the need for more data (since sample size and cell counts affect independence tests per Devore (1987), Fienberg (1977, 2007), Larntz (1978), MacDonald (2008) and Spirtes *et al.* (2000)), or the need for data that satisfies the underlying algorithmic assumptions. Spirtes *et al.* (2000) list these and other potential reasons (Spirtes *et al.*, 2000, p. 351). Each rationale could be further explored in order to provide additional insights; regardless, the previous analyses shows how the method identifies potential SIN structures, compares various structures and assists in gleaning SIN insights.

Regarding the output of FCI and the underlying algorithm, it appears that the imposition of constraints after the independence tests, may produce a structure that is not in accordance with the independence tests. This was observed in the structure produced when $\alpha = 0.1$, but does not appear to directly affect the portion of the structure examined in the hidden individual analysis provided in the next section.

6.4.3 Detecting Unrevealed Individuals

Upon examination of the FCI output with $\alpha = 0.05$ for the illustrative example, there are indications of latent variables, i.e. hidden individuals. For example, there is a bi-directed arc between Senators Reid and Cardin, and if the circle endpoint between Senators Harkin and Cardin represents an arrowhead, then another bi-directed arc exists. Consequently, it is possible a hidden individual could influence both Senators Reid and Cardin, while a different hidden individual could influence Senators Harkin and Cardin. Figure 12 depicts the SIN structure corresponding to these situations (note: graphical representation of latent and observed variable derived from Spirtes *et al.*, 2000). Such a causal structure is an elaboration of the causal structure on the right in Figure 11, and equivalent to the causal structure on the left in Figure 11.

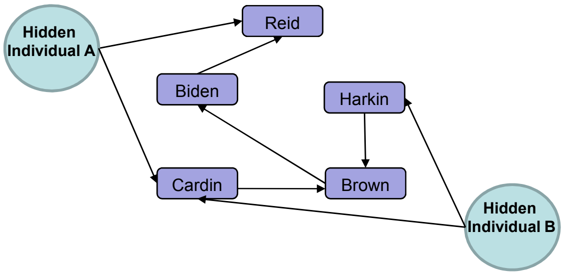


Figure 12 SIN with Hidden Individuals

Detecting an unrevealed individual was further examined using the approach where a node and its associated event data are removed from the sample, and subsequent exploratory analysis is performed. From the FCI output with $\alpha = 0.1$, the possibility exists, based on edge orientation, that the true SIN structure is one in which Senator Reid directly influences both Senators Biden and Cardin (cf. Figure 9). If the node and associated data for Senator Reid are removed, and subsequent causal exploratory analysis is performed, then it is reasonable to expect a hidden variable signature between Senators Biden and Cardin. Figures 13 and 14 provide the FCI output (with $\alpha = 0.05$ and $\alpha = 0.1$, respectively) for this case. The $\alpha = 0.05$ output with four Senators has an identical structure (less edge orientation) to what the original five Senator structure would be without the Senator Reid node and edges. The $\alpha = 0.1$ output contains an edge, with endpoint circles, between Senators Biden and Cardin. If the edge endpoints are replaced by arrowheads, the result is a double headed arrow. Based on the methodological assumptions, this reveals a hidden individual, possibly Senator Reid, who was not in the analyzed data. Additionally, the remainder of the SIN structure is nearly identical to the original.

The results from this section demonstrated the presented method's process for revealing a hidden individual; furthermore, follow-on analysis of an instance from the

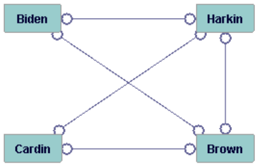


Figure 13 Causal Exploratory Results for Four Senators, $\alpha = 0.05$

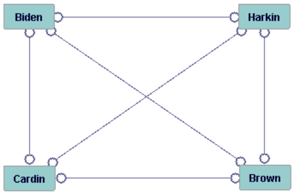


Figure 14 Causal Exploratory Results for Four Senators, $\alpha = 0.1$

real-world data indicated the method has the potential to reveal a hidden individual while reasonably maintaining the remainder of the SIN structure.

6.4.4 Causal Graph Reconstruction

As previously mentioned, the reconstruction method will be illustrated using connected, labeled DAGs with two directed edges, and there are twelve such graphs of order three (Korb & Nicholson, 2004; Robinson, 1977). The causal structure of a graph is given as the set of causal relations between any two nodes. A causal relation from labeled node x to labeled node y is denoted (xy), while the absence of a direct causal relation between labeled node x and labeled node y is denoted (x, y) (Shipley, 2002; West, 2001). The latter can alternatively be addressed by denoting only those causal relations that are present, and this notation is provided for DAG_u in Figure 15 (Spirtes *et al.*, 2000).

To illustrate the method for DAG_u, consider the situation (based upon the example of Figure 1 in Verma and Pearl (1991)) where dag_{t_1} is bc, i.e. the nodes b and c with a causal relation from b to c. This yields 4 possible, reconstructed causal graphs in $poss\_rec_{t_1}$ as shown in Figure 16. Consequently, $part\_acc_{t_1} = 0.25$. If subsequent causal exploratory analysis is performed for the three nodes (i.e.

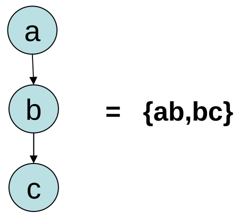


Figure 15 Unknown Causal Graph DAG_u

a, b, c) of DAG_u, then $pats$ consists of the three graphs in Figure 17. Consequently, $aug\_acc_{t_1} = |(poss\_rec_{t_1} \cap pats)|^{-1} = 0.5$.

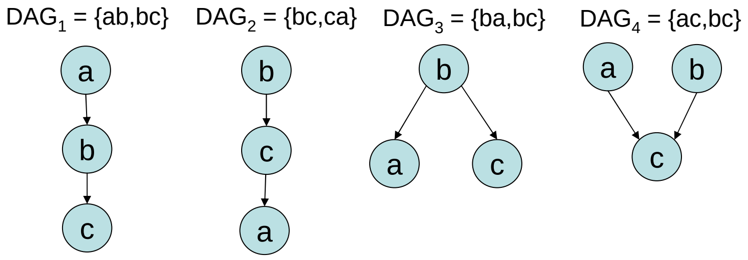


Figure 16 Possible Reconstructions Given dag_{t_1}

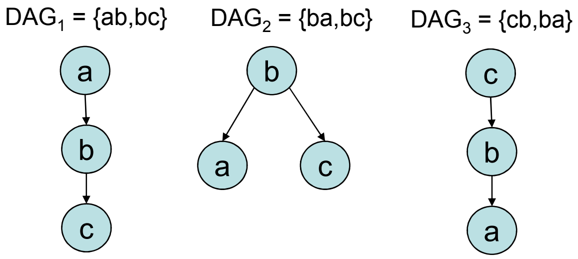


Figure 17 Equivalent Causal Graphs for the Three Nodes of DAG_u

If dag_{t_2} is a, c (i.e. the nodes a and c without a direct causal relation), then $poss\_rec_{t_2}$ is composed of the two graphs shown in Figure 18 and $part\_acc_{t_2} = 0.5$. Since a stable causal structure is assumed, $pats$ is unchanged when causal exploratory analysis is performed after t_2 instead of t_1; therefore, in this example, $aug\_acc_{t_1} = aug\_acc_{t_2} = 0.5$.

The remaining causal substructure, dag_{t_3}, is ab; therefore, $poss\_rec_{t_3} = DAG_u$ and $part\_acc_{t_3} = 1$. Consequently, subsequent causal exploratory analysis is not

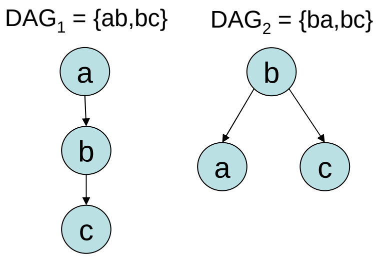


Figure 18 Possible Reconstructions Given dag_{t_2}

necessary/applicable. Tables 44 and 45 list $provseq_j, j = 1, 2, \ldots, 6$ and associated subgraphs, i.e. dag, for the twelve possible DAGs. The probability of choosing each of the twelve possible DAGs as the true causal graph, i.e. DAG_u, given the provided subgraphs dag is listed. Consequently, the probability of choosing $\{ab, bc\}$ (highlit in the tables; note this column in Table 45 is a duplicate of that in Table 44) as DAG_u (which it is indeed) is $part\_acc$. Additionally, those graphs which are causally equivalent to DAG_u, after the subgraph(s) have been provided, are given in the column $\{poss\_rec \cap pats\} - \{DAG_u\}$. From this column's information, aug_acc can be derived, and is reported in the final column.

Table 44 Causal Graph Reconstruction Details for DAG_u - Part A

provseq	*dag*	{ac,cb}	{ab,ac}	{ac,bc}	{ba,ac}	{bc,ca}	{ab,bc}
							part_acc
								{*poss_rec*∩*pats*}
								− {*DAG$_u$*}
									aug_acc

1	{bc}	0.0	0.0	0.25	0.0	0.25	0.25	{ba,bc}	0.5
	{a,c}	0.0	0.0	0.0	0.0	0.0	0.5	{ba,bc}	0.5
	{ab}	0.0	0.0	0.0	0.0	0.0	1.0	{}	1.0
2	{bc}	0.0	0.0	0.25	0.0	0.25	0.25	{ba,bc}	0.5
	{ab}	0.0	0.0	0.0	0.0	0.0	1.0	{}	1.0
	{a,c}	0.0	0.0	0.0	0.0	0.0	1.0	{}	1.0
3	{a,c}	0.0	0.0	0.0	0.0	0.0	0.25	{cb,ba};{ba,bc}	0.333
	{ab}	0.0	0.0	0.0	0.0	0.0	0.5	{}	1.0
	{bc}	0.0	0.0	0.0	0.0	0.0	1.0	{}	1.0
4	{a,c}	0.0	0.0	0.0	0.0	0.0	0.25	{cb,ba};{ba,bc}	0.333
	{bc}	0.0	0.0	0.0	0.0	0.0	0.5	{ba,bc}	0.5
	{ab}	0.0	0.0	0.0	0.0	0.0	1.0	{}	1.0
5	{ab}	0.0	0.25	0.0	0.0	0.0	0.25	{}	1.0
	{bc}	0.0	0.0	0.0	0.0	0.0	1.0	{}	1.0
	{a,c}	0.0	0.0	0.0	0.0	0.0	1.0	{}	1.0
6	{ab}	0.0	0.25	0.0	0.0	0.0	0.25	{}	1.0
	{a,c}	0.0	0.0	0.0	0.0	0.0	0.5	{}	1.0
	{bc}	0.0	0.0	0.0	0.0	0.0	1.0	{}	1.0

Table 45 Causal Graph Reconstruction Details for DAG_u - Part B

provseq	*dag*	{ba,bc}	{ab,cb}	{ca,ab}	{cb,ba}	{ca,cb}	{ca,ba}	{ab,bc}
								part_acc
									{*poss_rec*∩*pats*}
									− {*DAG$_u$*}
										aug_acc

1	{bc}	0.25	0.0	0.0	0.0	0.0	0.0	0.25	{ba,bc}	0.5
	{a,c}	0.5	0.0	0.0	0.0	0.0	0.0	0.5	{ba,bc}	0.5
	{ab}	0.0	0.0	0.0	0.0	0.0	0.0	1.0	{}	1.0
2	{bc}	0.25	0.0	0.0	0.0	0.0	0.0	0.25	{ba,bc}	0.5
	{ab}	0.0	0.0	0.0	0.0	0.0	0.0	1.0	{}	1.0
	{a,c}	0.0	0.0	0.0	0.0	0.0	0.0	1.0	{}	1.0
3	{a,c}	0.25	0.25	0.0	0.25	0.0	0.0	0.25	{cb,ba};{ba,bc}	0.333
	{ab}	0.0	0.5	0.0	0.0	0.0	0.0	0.5	{}	1.0
	{bc}	0.0	0.0	0.0	0.0	0.0	0.0	1.0	{}	1.0
4	{a,c}	0.25	0.25	0.0	0.25	0.0	0.0	0.25	{cb,ba};{ba,bc}	0.333
	{bc}	0.5	0.0	0.0	0.0	0.0	0.0	0.5	{ba,bc}	0.5
	{ab}	0.0	0.0	0.0	0.0	0.0	0.0	1.0	{}	1.0
5	{ab}	0.0	0.25	0.25	0.0	0.0	0.0	0.25	{}	1.0
	{bc}	0.0	0.0	0.0	0.0	0.0	0.0	1.0	{}	1.0
	{a,c}	0.0	0.0	0.0	0.0	0.0	0.0	1.0	{}	1.0
6	{ab}	0.0	0.25	0.25	0.0	0.0	0.0	0.25	{}	1.0
	{a,c}	0.0	0.5	0.0	0.0	0.0	0.0	0.5	{}	1.0
	{bc}	0.0	0.0	0.0	0.0	0.0	0.0	1.0	{}	1.0

As previously discussed, one can also examine the case where exploratory analysis is performed prior to receiving any causal substructure. As previously mentioned, there are twelve potential causal graphs for the considered situation; therefore, for DAG_u, $part\_acc_{t_0} = 0.083$. Performing exploratory analysis decreases the number of possibilities to three causally equivalent graphs; therefore, $aug\_acc_{t_0} = 0.333$. If causal substructure ab is provided at t_1, then the modified $poss\_rec_{t_1}$ yields an altered $part\_acc_{t_1} = aug\_acc_{t_1} = 1.0$, vice the original $part\_acc_{t_1} = 0.25$ as given in Tables 44 and 45. Additionally, as in the above analysis, subsequent causal substructure sequences can be coupled with the *a priori* exploratory analysis in order to determine the knowledge value of each substructure with respect to determining the true causal graph.

The results of this section show that the ability to reconstruct causal graphs may be enhanced by applying causal exploratory analysis. This is intuitively in keeping with the concept of incorporating prior knowledge into causal network development, which is possible in TETRAD. The contribution of the presented method is the derivation of a process that formalizes the integration of graph reconstruction with causal exploratory analysis resulting in associated accuracy metrics.

6.5 Conclusion

This chapter provided and illustrated a method to couple social influence with causal exploratory analysis to identify social influence network structures and reveal unknown individuals within such networks. The process for identifying a SIN and hidden individuals therein was demonstrated using real-world data. Additionally, a technique for assessing the analysis output against a proposed structure was presented. Given the assumptions, the method is a promising novel means to extract insights regarding influence in a social network. Additionally, an approach for coupling causal exploratory analysis and network reconstruction was introduced and demonstrated. Results showed that the ability to reconstruct causal graphs may be enhanced by applying causal exploratory analysis.

This chapter and the previous ones have presented methods to address the problem of characterizing and detecting unrevealed elements in network systems. While initial contributions have been made, further developments are possible. The following chapter provides some plausible areas for continued research.

7. Conclusions and Recommendations

7.1 Overview

Network systems are ubiquitous and can be very large. Additionally, characterizing and detecting unrevealed elements of network systems are truly difficult tasks; yet they are critically important in certain endeavors, e.g. counterterrorism. This dissertation makes steps toward addressing such systems and the attendant tasks. Techniques have been developed for application in extant problems; alternatively, new challenges have been raised and addressed in this dissertation, through a framework that is a generalization and expansion of a current problem. Additionally, given the underlying similarities (if not equivalence) of network representations, problems and problem solving techniques, the potential exists for interdisciplinary application of this dissertation's research (National Research Council, 2005; Newman, 2006). Research may raise more questions and uncovers new challenges with ideas for addressing them (C. Moorman, personal communication, October 14, 2008). Consequently, not only are contributions addressed in this chapter, but also recommendations for future research.

7.2 Contributions

The crux of this dissertation was the development and demonstration of methods to characterize and detect unrevealed elements in network systems. The associated contributions were three-fold:

1. A method to identify (and consequently characterize) and detect individuals that bridge groups in social networks. In network terms, these individuals were referred to as connection nodes.

2. A method to reconstruct network structures given repeat observations of various parts of the network structure. Additionally, an approach to address the reconstruction of causal (or influence) networks was provided.
3. A method to identify potential social influence network (SIN) structures and hidden individuals using causality analysis.

With respect to the first contribution, the research demonstrated the feasibility of a method developed for revealing bridge elements in social networks. The method was tested on both empirical and generated data sets. A technique for where to place the revealed bridges was also provided and demonstrated. Evident, from the analysis associated with this contribution, was the fact that in order for the method to provide reasonable results (i.e. high classification accuracy and/or low false positive rates) there must be attributes that distinguish a bridge from a non-bridge. This issue will be addressed in the section on future research; however, as presented the method provides an initial capability to not only classify, but also detect key network elements, i.e. connection nodes.

Pursuant to the first contribution, it was noted that there may exist conditions under which observations of partial network structures can result in the revelation of other network elements. The second contribution addressed a related problem; specifically, the determination of a network structure from (possibly) repeated observations of a proper subset of the network structure. A new framework for reconstructing a network from its subnetworks, a general formula for reconstruction accuracy within the framework and a demonstration of the results of the associated analysis were provided. Traditional reconstruction poses reconstruction in terms of the number of subgraphs required to reconstruct the original network. This research associates that number, in a time context, i.e. the duration to reconstruct, since repeat observations are permissible. Additionally, stochasticity is addressed via the definition of accuracy introduced in the modified framework. Both accuracy and duration are affected by the sequence of observations, as well as, the set of non-isomorphic vertex deleted subgraphs. Consequently, the framework provides an approach to assessing possible network structures. Furthermore, insight into potential mis-identification can be derived from the associated analysis. Ultimately, the modified framework is a more general formulation for network reconstruction, and can address an, arguably, more difficult reconstruction problem.

The final contribution provided and illustrated a method to couple social influence with causal exploratory analysis to identify SIN structures and reveal unknown individuals within such networks. The process for identifying a SIN and hidden individuals therein was demonstrated using real-world data. Given the assumptions, the method showed promise as means to gain insights regarding both structure and membership of a SIN. Such knowledge is valuable when trying to understand networks and associated various courses of actions with respect to the structure, e.g. who is (should) fulfilling (fulfill) certain organizational roles (e.g. Allen, 1977; Conway, 1997; Lawrence & Lorsch, 1967; Schwartz & Jacobson, 1977). Additionally, a method that formalizes the integration of graph reconstruction with causal exploratory analysis and provides associated accuracy metrics was introduced and demonstrated. Results showed that the ability to reconstruct causal graphs may be enhanced by applying causal exploratory analysis.

7.3 Future Research

As previously mentioned, concepts in this dissertation should be applicable to network types other than those demonstrated. Consequently, the bridge identification and detection method should be feasible for other than social networks (to include those with directed edges where relevant centrality measures, such as those in Wasserman and Faust (1994), could be employed); assuming there are distinct connection nodes. In fact, data from a physical system network would likely contain less variation; therefore, improved classification results are expected with such networks. Not only could different network types be investigated, but also additional network structures and node attributes. This could be accomplished by generating networks through a design of experiment approach. Another area of research is the method whereby revealed bridges are inserted. One possible method is to combine the presented insertion methods with graph reconstruction concepts and social influence analysis. Such an approach may provide a more macro-level solution to address not only connection nodes, but also other network elements pursuant to the overarching goal of this dissertation. The following paragraphs provide potential contexts and proposed approaches for follow-on analyses related to causal networks (of which social influence networks are a proper subset) and/or graph reconstruction.

While meaningful causal analysis can be performed when the network nodes are known, i.e. labeled, there may be circumstances where the number of nodes is known but associated identities or labels are not, e.g. very noisy observations of activities, individuals or objects under surveillance (analogous to link noise of Kubica *et al.* (2003b)). In such situations where the objective is network reconstruction, it is helpful to deal with networks represented graphically by a structure other than DAGs, even though the graphical representation of causal networks are DAGs. Such a choice is useful for making initial gains in this area, since not every DAG is reconstructable (Stockmeyer, 1981, pp. 234-235, 237). Consequently, a reconstruction method could employ a DAG (denoted DAG_u) approximation such as an oriented tree, which is a tree whose edges have been made directed. In such analysis, DAG_u could be denoted as OT_u. The utility of this approximation is that an oriented tree with three or more end points (nodes having degree one) can be reconstructed from its unlabeled vertex-deleted (oriented) subtrees (Harary & Palmer, 1966, pp. 803, 809-810).

The following approach addresses the previous case involving unlabeled, causal graphs represented by oriented trees. The associated methodology assumes a stable causal structure exists; however, since the vertex-deleted subgraphs are unlabeled, the observer must perform exploratory analysis for multiple subgraphs, possibly incurring redundancy and introducing additional uncertainty into the reconstruction process. For example, Spirtes *et al.* (1990) provided analysis based upon Rogers and Maranto (1989) that incorrectly omitted edges in the causal structure. Such an error could be rectified, but it could also be perpetuated yielding uncertainty regarding the true structure (Shipley, 2002, pp. 71, 278-280; Spirtes *et al.*, 1990, pp. 185-187, 189, 191, 197; Spirtes *et al.*, 2000, pp. 77, 82, 84, 93, 95-96).

1. Assume there exists a ground truth causal network structure, OT_u, not known to an observer; however, the number of network nodes, n, is known.
2. At time $t = 0$, $n - 1$ nodes of OT_u are (probabilistically) generated.
3. Exploratory analysis is performed on the $n - 1$ nodes.
4. At each time step, $t_i, i \in \mathbb{Z}^+$, $n - 1$ nodes of OT_u are chosen according to some probabilistic decision rule.
5. Exploratory analysis is performed on the $n - 1$ nodes.
6. Continue node generation and exploratory analysis until OT_u has been reconstructed or a desired number of time steps has been completed.

Once this process has been performed, the following questions can be addressed.

1. For different probabilistic decision rules, what is the mean number of observations (and variance) at which the entire network is reconstructed, if at all?

   Harary and Palmer (1966) proved that oriented trees containing at least three vertices of degree one can be reconstructed from the oriented vertex deleted subtrees, i.e. not every vertex deleted subgraph is required. Their result is a possible lower bound. Furthermore, due to issues such as causal model equivalence and inherent uncertainty in the exploratory analysis process, the result of each time step's causal exploratory analysis may be incorrect or not entirely revealing (e.g. a pattern containing undirected and directed edges, or a partially oriented inducing path graph) (Shipley, 2002, pp. 71, 275, 278-280, 282; Verma & Pearl, 1991, pp. 259-260). The analysis of Spirtes *et al.* (1990, 2000) based upon Rogers and Maranto (1989) serves to illustrate this point (Spirtes *et al.*, 1990, pp. 185-187, 189, 191, 197; Spirtes *et al.*, 2000, pp. 6, 59, 61, 77, 82-87, 93, 95-96, 139-140).

2. What graph characteristics, processes and probabilistic mechanisms, if any, impact the duration required to reconstruct the original network, and why (e.g. McMullen, 2005)?

   A partial answer to this question may involve properties and methods of causal exploratory analysis. For example, sampling error during the causal exploratory process could result in one set of $n - 1$ nodes not containing the true (but unknown to the observer) edges; therefore, depending on the probabilistic mechanism employed, reconstruction may require several extra time steps (Shipley, 2002, pp. 247-248, 275, 278-279, 282). The analysis by Spirtes *et al.* (1990) again provides a sample scenario that could result in additional time steps. (Spirtes *et al.*, 1990, pp. 184-186, 189, 191; Spirtes *et al.* 2000, pp. 82-83)

3. What graph characteristics and probabilistic mechanisms impact the accuracy of reconstructing the original network, and why?

   As previously discussed, the reconstructed graph may contain uncertainty due to the exploratory process. Some of the uncertainty is also influenced by certain features of the ground truth network, e.g. links that point to the same vertex (Verma & Pearl, 1991, pp. 259-260, 264). The impact of these features on the reconstruction process, e.g. the number of such links contained in the ground truth network, could be examined. The number of observations, i.e. time steps, can also affect the reconstruction accuracy; for which it is reasonable to use the number of equivalent causal models resulting from the reconstruction process as an associated measure.

Another analytic excursion involving unlabeled graphs is the situation where DAG_u does not necessarily represent a causal network. For this network type, it is plausible that an analyst would be interested in the overall structure without initial concern of or access to node labels. For example, a social network analyst may be interested in the structure of a group of individuals apart from the identity of the individuals (Wasserman & Faust, 1994, pp. 419-420, 423; White, Boorman, & Breiger, 1976, pp. 731, 742, 744). Again, in order to uniquely reconstruct the network, the examined networks could be represented graphically by oriented trees with at least three end points.

Furthermore, an approach analogous to the previous process could be employed to address the case involving unlabeled, oriented trees representing non-causal networks. The key differences would be the manner in which subgraphs are provided and the omission of causal exploratory analysis. The provision of subgraphs and associated analysis could be similar to the method presented for unlabeled, undirected graphs. That method did not constrain the (probabilistic) mechanism in which subgraphs were observed; however, for both unlabeled, oriented trees representing non-causal networks and unlabeled, undirected graphs, the impact of assuming a particular observation scheme could be analyzed.

An additional excursion for not only unlabeled, oriented trees representing non-causal networks, but also unlabeled, undirected graphs involves uncertainty regarding the existence of edges. For example, if at each time step, t_i, the probability of edge detection, p_d, is not one, then uncertainty regarding the true structure of the unknown graph may result. A simplifying assumption could be implemented, e.g. one never detects an edge that is not actually present in the unknown graph, i.e. no false positives.

In an attempt to broadly apply the concepts in this dissertation, research regarding networks that contain both directed and undirected relations would also be merited. Such networks could be portrayed as mixed multigraphs without loops, and vertices could be labeled or unlabeled (Gardner, Bobga, Nguyen, & Coker, 2005, p. 13; Harary, 1969, p. 10; Wasserman & Faust, 1994, pp. 73-75; West, 2001, p. xvi). Depending on the application, the relations could be causal, non-causal or a mix of the two. A causal interpretation would be appropriate for addressing the notions of multiple "... influence-related activities ..." (March, 1955, p. 436) or spheres of influence (March, 1949, p. 213). Such a representation would, according to March (1955), be more indicative of the influence relationship between individuals than the single relation approach. Identifying possible SIN structures and revealing hidden individuals could potentially be accomplished by extrapolating the methods in this research. Additionally, the aforementioned graph reconstruction excursions (e.g. repeat observations) for the previously mentioned graph types could be analyzed for mixed multi-graphs; however, different techniques, e.g. reconstructability analysis, may be required (Klir, 1985, pp. 222-223; Klir & Parviz, 1986; Zwick, 2004, p. 889). Beyond the multiple relation network is the network of networks representation. Such a structure could contain, m (possibly multiple relation) social networks that are interconnected. While discussion, representation and research of multiple relation networks and network of networks have occurred (e.g. Hamill, 2006, pp. 5-6; Kennedy, 2003, pp. 3-1, 3-2; National Research Council, 2005, pp. 7; Renfro, 2001, pp. 6, 108; Wasserman & Faust, 1994, pp. 73-76), there appears to be an open area of research regarding SIN structure identification and detection of hidden individuals.

As mentioned in Chapter 6, there appear to be inconsistencies within the FCI algorithm for producing causal structures. This provides an area for further investigation, so there is a clear understanding of limitations associated with interpreting and applying FCI results, and to propose corrective measures.

7.4 Conclusion

Researchers from a variety of fields have examined hidden nodes and links in different contexts. Accordingly, they have developed viable techniques for addressing portions of the unrevealed elements problem. Through synthesizing extant concepts and previous contributions, this dissertation extends the body of knowledge in this area. Networks are based on relationships, wherein lie both problems and solutions. This dissertation offers a means to understand the former, in order to provide the latter.

Bibliography

1. Adafre, S. F., & de Rijke, M. (2005) Discovering missing links in Wikipedia. In R. Grossman, R. J. Bayardo, & K. P. Bennett (Eds.), *Proceedings of the 3rd International Workshop on Link Discovery, Conference on Knowledge Discovery in Data, LinkKDD05* (pp. 90-97). New York: ACM Press.

2. Al Hasan, M., Chaoji, V., Salem, S., & Zaki, M. J. (2006). Link prediction using supervised learning. In *The Proceedings of the Fourth Workshop on Link Analysis, Counterterrorism and Security (held in conjunction with: Sixth SIAM International Conference on Data Mining (SDM 06)).* Retrieved March 12, 2007, from http://www.siam.org/meetings/sdm06/workproceed/Link%20Analysis/12.pdf

3. Allen, T. J. (1977) *Managing the flow of technology: Technology transfer and the dissemination of technological information within the r&d organization.* Cambridge, MA: MIT Press.

4. Association for the Advancement of Artificial Intelligence (1998). Link analysis background webpage regarding the 1998 AAAI fall symposium on artificial intelligence and link analysis. Retrieved February 22, 2007, from http://kdl.cs.umass.edu/events/aila1998/link-analysis.html.

5. Baldwin, J. (2003). *Graph reconstruction numbers.* Master's project report, Rochester Institute of Technology, Rochester, New York. Retrieved November 3, 2007, from https://ritdml.rit.edu/dspace/bitstream/1850/2745/2/JBaldwin MasterProject2003.pdf

6. Banks, D. and Carley, K. (1994). Metric inference for social networks. *Journal of Classification, 11*, 121-149.

7. Batt, T. (2002, September 8). Reid firmly rooted in Mormon faith. *Las Vegas Review-Journal.* Retrieved from http://www.reviewjournal.com/lvrj_home/2002/Sep-08-Sun-2002/news/19525007.html

8. Baumes, J., Goldberg, M., Madgon-Ismail, M., & Wallace, A. (2004). Discovering hidden groups in communications networks. In H. Chen, R. Moore, D. D. Zeng, & J. Leavitt (Eds.), *Intelligence and Security Informatics, Second Symposium on Intelligence and Security Informatics, ISI 2004. Lecture Notes in Computer Science 3073* (pp. 378-389). New York: Springer.

9. Berlo, D. K., Lemert, J. B., & Mertz, R. J. (1969). Dimensions for evaluating the acceptability of message sources. *Public Opinion Quarterly, 33*, 563-576.

10. Binder, J., Koller, D., Russell, S., & Kanazawa, K. (1997). Adaptive probabilistic networks with hidden variables. *Machine Learning, 29*, 213-244.

11. Blizard, W. D. (1989). Multiset theory. *Notre Dame Journal of Formal Logic, 30*(1), 36-66.

12. Bogart, K. P. (1983). *Introductory combinatorics.* Boston: Pitman.

13. Bollobás, B. (1990). Almost every graph has reconstruction number three. *Journal of Graph Theory, 14*(1), 1-4.

14. Bollobás, B. (2001). *Random Graphs* (2nd ed.). Cambridge, England: Cambridge University Press.

15. Bondy, J. A. (1978). Reflections on the legitimate deck problem. In D. A. Holton & J. Seberry (Eds.), *Combinatorial mathematics*, (pp. 1-12). Berlin: Springer-Verlag.

16. Bondy, J. A. (1991). A Graph reconstructor's manual. In A. D. Keedwell (Ed.), *London mathematical society lecture note series, 166: Surveys in combinatorics, 1991* (pp. 221-252). Cambridge, England: Cambridge University Press.

17. Bondy, J. A., & Hemminger, R. L. (1977). Graph reconstruction - A survey, *Journal of Graph Theory, 1*, 227-268.

18. Borgatti, S. P., Carley, K. M., & Krackhardt, D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social Networks, 28*, 124-136.

19. Boyen, X., Friedman, N., & Koller, D. (1999). Discovering the hidden structure of complex dynamic systems. In K. B. Laskey, & H. Prade (Eds.), *Proceedings of the 15th Annual Conference on Uncertainty in AI (UAI)* (pp. 91-100). San Francisco: Morgan Kauffman.

20. Breese, J. S., Heckerman, D. & Kadie, C. (1998). *Empirical analysis of predictive algorithms for collaborative filtering.* (Technical Report, MSR-TR-98-12; May, 1988; revised October, 1988). Redmond, WA: Microsoft Research, Microsoft Corporation.

21. Bryant, R. M. (1971). On a conjecture concerning the reconstruction of graphs. *Journal of Combinatorial Theory Series B, 11*, 139141.

22. Bullock, C.S. III, & Brady, D. W. (1983). Party, constituency, and roll-call voting in the U.S. Senate. *Legislative Studies Quarterly, 8*(1), 29-43.

23. Burt, R. S. (1987). A note on missing network data in the general social survey. *Social Networks, 9*(1), 63-73.

24. Butts, C. T. (2003). Network inference, error, and informant (in)accuracy: A Bayesian approach. *Social Networks, 25*(2), 103-140.

25. Cai, D., Shao, Z., He, K., Yan, X., & Han, J. (2005). Mining hidden community in heterogeneous social networks. In R. Grossman, R. J. Bayardo, & K. P.

Bennett (Eds.), *Proceedings of the 3rd International Workshop on Link Discovery, Conference on Knowledge Discovery in Data, LinkKDD05* (pp. 58-65). New York: ACM Press.

26. Carley, K. M., Dombroski, M., Tsvetovat, M., Reminga, J., & Kamneva, N. (2003). Destabilizing dynamic covert networks. In *Proceedings of the 8th International Command and Control Research and Technology Symposium.* Retrieved November 17, 2000, from http://www.dodccrp.org/events/8th_ICCRTS/pdf/021.pdf. Washington, DC: DoD Command and Control Research Program.

27. Cavallo, R.E. (1980). Reconstructability and identifiability in the evaluation of structure hypotheses: An issue in the logic of modelling. In B. Banathy (Ed.), *Systems Science and Science* (pp. 647-654). Louisville, KY: Society for General Systems Research.

28. Cavallo, R. E., & Klir, G. J. (1979). Reconstructability analysis of multidimensional relations: A theoretical basis for computer-aided determination of acceptable systems models. *International Journal of General Systems, 5*(3), 143-171.

29. Cavallo, R. E., & Klir, G. J. (1981). Reconstructability analysis: Overview and bibliography. *International Journal of General Systems, 7*(1), 1-6.

30. Chemical Rubber Company (1996). *CRC standard mathematical tables and formulae.* Boca Raton, FL: CRC Press.

31. Clark, C. R. (2005). *Modeling and analysis of clandestine networks.* Master's thesis, Air Force Institute of Technology, Wright-Patterson AFB, Ohio. Retrieved November 17, 2008, from http://handle.dtic.mil/100.2/ADA439611

32. Combating Terrorism Center (2006). *Letter Exposes New Leader in Al-Qa'ida High Command.* A Combating Terrorism Center report. Retrieved September, 29, 2006, from http://ctc.usma.edu/publications/pdf/CTC-AtiyahLetter.pdf.

33. Cohn, D., & Hofmann, T. (2001). The missing link - A probabilistic model of document content and hypertext connectivity. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference* (pp. 430-436). Cambridge, MA: MIT Press.

34. Connolly, D. (1993). Constructing hidden variables in Bayesian networks via conceptual learning. In *Machine Learning, Proceedings of the Tenth International Conference* (pp. 65-72). San Mateo, CA: Morgan Kaufmann.

35. Conway, S. (1997). Strategic personal links in successful innovation: Link-pins, bridgess and liaisons. *Creativity and Innovation Management, 6*(4), 226-233.

36. Cooke, R. J. E. (2006). *Link prediction and link detection in sequences of large social networks using temporal and local metrics.* Master's thesis, University of Cape Town, Cape Town, South Africa. Retrieved November 17, 2008, from http://pubs.cs.uct.ac.za/archive/00000370/01/Richard_Cooke_-_Link_prediction_and_link_detection_in_sequences_of_large_social_networks_using_temporal_and_local_metrics_(masters_dissertation)_-_2006.pdf

37. Cooper, G. F. (1999). An overview of the representation and discovery of causal relationships using Bayesian networks. In C. Glymour, & G. F. Cooper (Eds.). *Computation, causation and discovery* (pp. 1-62). Menlo Park, CA: AAAI Press.

38. Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*, 309-347.

39. Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). *Introduction to algorithms* (2nd ed.). Cambridge, MA: MIT Press.

40. Costenbader, E., & Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social Networks, 25*(3), 283-307.

41. Dean, T., & Kanazawaw, T. (1989). A model for reasoning about persistence and causation. *Computational Intelligence, 5*, 142-150.

42. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1-38.

43. Devore, J. L. (1987). *Probability and statistics for engineering and the sciences* (2nd ed.). Montery, CA: Brooks/Cole Publishing Company.

44. Dillon, W. R., & Goldstein, M. (1984). *Multivariate analysis. Methods and application.* New York: John Wiley and Sons, Inc.

45. Domingos, P., & Richardson, M. (2001). Mining the network value of customers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)* (pp. 57-66). New York: ACM Press.

46. Doreian, P. (2001). Causality in social network analysis. *Sociological Methods & Research, 30*(1), 81-114.

47. Doyle, R. J. (1989). Reasoning about hidden mechanisms. In N. S. Sridharan (Ed.), *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence* (pp. 1343-1349). San Mateo, CA: Morgan Kaufmann.

48. Duncan, O. D., Haller, A. O., & Portes, A. Peer influences on aspirations: A reinterpretation. *The American Journal of Sociology, 74*(2), 119-137.

49. Eisenstadt, S. N. (1952). Communication processes among immigrants in Israel. *Public Opinion Quarterly, 16*, 42-58.

50. Elidan, G., Lotner, N., Friedman, N., & Koller, D. (2000). Discovering hidden variables: A structure-based approach. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000* (pp. 479-485). Cambridge, MA: MIT Press

51. Englander, I. (2003). *The architecture of computer hardware and systems software: An information technology approach* (3rd ed.). New York: Wiley.

52. Epstein, A. L. (1961). The network and urban social organization. *Rhodes-Livingstone Journal, 29*, 29-62.

53. Erdős, P., & Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute, Hungarian Academy of Sciences, 5*, 17-61.

54. Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). London: Arnold.

55. Ferrand, A., Mounier, L., & Degenne, A. (1999). The diversity of personal networks in France: Social stratification and relational structures. In B. Wellman (Ed.), *Networks in the global village: A life in contemporary comnunities* (pp. 185-224). Boulder, CO: Westview Press.

56. Fienberg, S. E. (1977). *The analysis of cross-classified categorical data.* Cambridge, MA: MIT Press.

57. Fienberg, S. E. (2007). *The analysis of cross-classified categorical data* (paperback ed.). New York: Springer-Verlag.

58. Flake, G. W., Lawrence, S., & Giles, C. L. (2000). Efficient identification of web communities. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 150-160). New York: ACM Press. Retrieved from http://www.cs.washington.edu/education/courses/cse522/05au/communities-flake.pdf

59. Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association, 81*, 832-842.

60. Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry, 40*(1), 35-41.

61. Freeman, L. C. (1980). The gatekeeper, pair-dependency and structural centrality. *Quality and Quantity, 14*, 585-592.

62. French, J. R. P (1956). A formal theory of social power. *Psychological Review, 63*(3), 181-194.

63. Friedkin, N. E. (1986). A formal theory of social power. *Journal of Mathematical Sociology, 12*(2), 103-126.

64. Friedkin, N. E. (1990). Social networks in structural equation models. *Social Psychology Quarterly, 53*(4), 316-328.

65. Friedkin, N. E. (1998). A structural theory of social influence. Cambridge, England: Cambridge University Press.

66. Friedman, N. (1997). Learning belief networks in the presence of missing values and hidden variables. In D. Fisher (Ed.), *Machine learning: Proceedings of the Fourteenth International Conference* (pp. 125-133). San Francisco: Morgan Kaufmann.

67. Friedman, N., Murphy, K., & Russell. S. (1998). Learning the structure of dynamic probabilistic networks. In G. F. Cooper, & S. Moral (Eds.), *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI)* (pp. 139-147). San Francisco: Morgan Kaufmann.

68. Gardner, R., Bobga, B., Nguyen, C., & Coker, G. (2005). Some graph, digraph, and mixed graph results concerning decompositions, packings, and coverings. Presented at the Joint Meeting of the A.M.S. and the M.A.A., *AMS Special Session on Design Theory and Graph Theory, I*, Atlanta, Georgia, January 5, 2005. Retrieved March 27, 2007, from http://www.etsu.edu/math/gardner/talks/ams05.pdf.

69. Garson, G. D. (n.d.). Logistic regression. In *Statnotes: Topics in multivariate analysis.* Retrieved December 5, 2007, from http://www2.chass.ncsu.edu/garson/pa765/statnote.htm.

70. Geiger, D., & Pearl, J. (1990). On the logic of causal models. In L. Kanal, T. Levitt, R. Shachter, & J. F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence 4*, (pp. 3-14). Amsterdam: Elsevier Science Publishers B. V. (North-Holland).

71. Getoor, L., Friedman, N., Koller, D., & Taskar, B. (2002). Learning probabilistic models of link structure [Electronic version]. *Journal of Machine Learning Research, 3*, 679-707.

72. Gibson, D., Kleinberg, J., & Raghavan, P. (1998). Inferring web communities from link topology. In R. Akscyn (Ed.) *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia* (pp. 225-234). New York: ACM.

73. Gillham, P. F. & Marx, G. T. (2000). Complexity & irony in policing and protesting: The World Trade Organization in Seattle. *Social Justice, 27*(2), 212-236.

74. Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling.* Orlando, FL: Academic Press.

75. Glymour, C., Scheines, R., Spirtes, P., & Ramsey, J. (2004a). *TETRAD IV manual.* Retrieved from http://www.phil.cmu.edu/projects/tetrad_download/files/manual.pdf

76. Glymour, C., Scheines, R., Spirtes, P., & Ramsey, J. (2004b). TETRAD IV (Version 4.3.9-0) [Software]. Available from http://www.phil.cmu.edu/projects/tetrad_download/files/manual.pdf

77. Glymour, C., & Spirtes, P. (1988). Latent variables, causal models and overidentifying constraints. *Journal of Econometrics, 39*(1-2), 175-198.

78. Goldenberg, J., Libai, B., & Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters, 12*(3), 211-223. Retrieved February 21, 2007, from ABI/INFORM Research database.

79. Goldenberg, A., Kubica, J., & Komarek, P. (2003). A Comparison of statistical and machine learning algorithms on the task of link completion. In *Proceedings of the Workshop on Link Analysis for Detecting Complex Behavior (LinkKDD2003).* New York: ACM. Retrieved June 6, 2006, from http://www-2.cs.cmu.edu/ dunja/LinkKDD2003/papers/Goldenberg.ps

80. Goldhamer, H., & Shils, E. A. (1939). Types of power and status. *The American Journal of Sociology, 45*(2), 171-182.

81. Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I - A modified latent structure approach. *American Journal of Sociology, 7*(5), 1179-1259.

82. Gower, J. C. (1971). A general coefficient of similiarity and some of its properties. *Biometrics, 27*, 857-874.

83. Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology, 78*(6), 1360-1380.

84. Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through blogspace [Electronic version]. In *Proceedings of the 13th International World Wide Web Conference (WWW04)* (pp. 491-501). New York: ACM. Retrieved February 26, 2007, from http://portal.acm.org/ft_gateway.cfm?id=988739&type=pdf&coll=GUIDE&dl =GUIDE&CFID=11932451&CFTOKEN=64928309

85. Haimes, Y. Y. (2004). *Risk modeling, assessment, and management.* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

86. Haller, A. O., & Butterworth, C. E. (1960). Peer influences on levels of occupational and educational aspiration. *Social Forces, 38*(4), 289-295.

87. Hamill, J. T. (2006). *Analysis of layered social networks* (Doctoral dissertation, Air Force Institute of Technology, 2006). Retrieved from http://www.dtic.mil.

88. Hammer, M. (1979/1980). Predictability of social connections over time. *Social Networks, 2*, 165-180.

89. Harary, F. (1964). On the reconstruction of a graph from a collection of subgraphs. In M. Fiedler (Ed.), *Theory of graphs and its applications* (pp. 47-52). New York: Academic.

90. Harary, F. (1969). *Graph theory.* Reading, MA: Addison-Wesley Publishing Company.

91. Harary, F., & Manvel, B. (1970). The reconstruction conjecture for labeled graphs. In R. K. Guy, H. Hanani, N. Sauer, & J Schőnheim (Eds.), *Combinatorial structures and their applications* (Proceedings of the Calgary International Conference on Combinatorial Structures and Their Applications, Calgary, Alberta, 1969), (pp. 131-146). New York: Gordon and Breach.

92. Harary, F., & Norman, R. Z. (1953). *Graph theory as a mathematical model in social science.* Ann Arbor, MI: University of Michigan Press.

93. Harary, F., & Palmer, E. (1966). The reconstruction of a tree from its maximal subtrees. *Canadian Journal of Mathematics, 18*, 803-810.

94. Harary, F., & Plantholt, M. (1985). The graph reconstruction number. *Journal of Graph Theory, 9*, 451-454.

95. Harper, W. R., & Harris, D. G. (1975). The application of link analysis to police intelligence. *Human Factors, 17*(2), 157-164.

96. Heise, D. R. (1975). *Causal analysis.* New York: John Wiley and Sons.

97. Hemaspaandra, E., Hemaspaandra, L. A., Radziszowski, S. P., & Tripathi, R. (2007). Complexity results in graph reconstruction. *Discrete Applied Mathematics, 155*, 103118.

98. Hickman, J. L. (1980). A note on the concept of multiset. *Bulletin of the Australian Mathematical Society, 22*, 211217.

99. Hoff, P. D., & Ward, M. D. (2006). *VMASC statistics and social network analysis project report* (October 9, 2006).

100. Horsfall, A. B. , & Arensberg, C. M. (1949), Teamwork and productivity in a shoe factory. *Human Organization, 8*(Winter), 13-25.

101. Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

102. Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion: Pyschological studies of opinion change.* New Haven: Yale University Press.

103. Hovland, C. I., & Weiss, W. (1951), The influence of source credibility on communication effectiveness. *Public Opinion Quarterly, 15*, 635-660.

104. Howard, R. A. (1988). Decision analysis: Practice and Promise. *Management Science, 34*(6), 679-695.

105. Hunter, D. R., Goodreau, S. M., & Handcock, M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association, 103*(481), 248-258. doi:10.1198/016214507000000446

106. JMP 6.0.0 (2005). [Computer Software]. Cary, NC: SAS Institute.

107. Jacobson, E., & Seashore, S. E. (1951). Communication practices in complex organizations. *Journal of Social Issues, 7*, 28-40.

108. Jackson, R. H. F., Boggs, P. T., Nash, S. G., & Powell, S. (1991). Guidelines for reporting results of computational experiments. Report of the ad hoc committee. *Mathematical Programming, 49*, 413-425.

109. James, L. R. , Muliak, S. A., & Brett, J. M., (1982). *Causal analysis. Assumptions, models and data.* Beverly Hills: Sage Publications.

110. Jensen, F.V. (2001). *Bayesian networks and decsion graphs.* New York: Springer-Verlag, New York, Inc.

111. Ji, X., & Zha, H. (2004). Sensor positioning in wireless ad-hoc sensor networks using multidimensional scaling. In *Proceedings of IEEE INFOCOM 2004. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies, 4* (pp. 2652-2661). Piscataway, NJ: IEEE.

112. Jöreskog, K. G., & Sörbom, D. (1995) *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Lincolnwood, IL: Scientific Software International.

113. Katz, E., & Lazarsfeld, P. F. (1955). *Personal influence: The part played by people in the flow of mass communications.* Glencoe, IL: The Free Press.

114. Katz, E., & Lazarsfeld, P. F. (1964). *Personal influence: The part played by people in the flow of mass communications* (1st pbk. ed.). Glencoe, IL: The Free Press.

115. Kelly, P. J. (1957). A congruence theorem for trees. *Pacific Journal of Mathematics, 7*(1), 961-968.

116. Kemeny, J. G. (1959). Mathematics without numbers. *Daedalus, 88*, 577-591.

117. Kempe, D., Kleinberg, J., & Tardos, E. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining (KDD'03)* (pp. 137-146). New York: ACM Press.

118. Kennedy, K. T. (2003) *An analysis of multiple layered networks.* Master's thesis, Air Force Institute of Technology, Wright-Patterson AFB, Ohio. Retrieved November 26, 2008, from http://handle.dtic.mil/100.2/ADA420865

119. Killworth, P. D., & Bernard, H. R. (1976). Informant accuracy in social network data. *Human Organization, 35*(3), 269-286.

120. Kjærulff, U. (1995). dHugin: A computational system for dynamic time-sliced Bayesian networks. *International Journal of Forecasting, 11*, 89-111.

121. Kleinbaum, D. G., & Klein, M. (2002). *Logistic regression: A self-learning text* (2nd ed.). New York: Springer-Verlag Inc.

122. Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the Association for Computing Machinery, 46*(5), 604-632.

123. Klir, G. J. (1985). *Architecture of systems problem solving.* New York: Plenum Press.

124. Klir, G. J., & Parviz, B. (1986). General reconstruction characteristics of probabilistic and possibilistic systems. *International Journal of Man-Machine Studies, 25*, 367-397.

125. Komarek, P. (2004). *Logistic regression for data mining and high-dimensional classification* (Technical Report, 04-34, May, 2004) Pittsburgh, PA: Robotics Institute, Carnegie Mellon University.

126. Korb, K. B., & Nicholson, A. E. (2004). *Bayesian artificial intelligence.* Boca Raton, FL: Chapman and Hall/CRC.

127. Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks, 28*, 247-268.

128. Krackhardt, D. (1987). Cognitive social structures. *Social Networks, 9*, 109-134.

129. Krebs, V. E. (2002). Mapping networks of terrorist cells. *Connections, 24*(3), 43-52.

130. Kubica J., Moore, A., Schneider, J., & Yang, Y. (2002). Stochastic link and group detection. In R. Dechter, M. Kearns, & R. Sutton (Eds.), *Proceedings of the 18th National Conference on Artificial Intelligence* (pp. 798-804). Menlo Park, CA: AAAI Press.

131. Kubica J., Moore A., & Schneider, J. (2003). Tractable group detection on large link data sets. In *ICDM, Third IEEE International Conference on Data Mining* (pp. 573-576). Washington, DC: IEEE Computer Society.

132. Kubica, J. M., Moore, A., Cohn, D., & Schneider, J. (2003). Finding underlying connections: A fast graph-based method for link analysis and collaboration queries. In *Proceedings of the Twentieth International Conference on Machine Learning, ICML 2003* (pp. 392-399). Menlo Park, CA: AAAI Press.

133. Kushnir, T., Gopnik, A., Schulz, L., & Danks, D. (2003). Inferring hidden causes. In R. Alterman, & D. Kirsh (Eds.), *Proceedings of the 25th Conference of the Cognitive Science Society.*

134. Lachenbruch, P. A., Sneeringer, C., & Revo, L. T. (1973). Robustness of the linear and quadratic discriminant function to certain types of non-normality. *Communications in Statistics - Theory and Methods, 1*(1), 39-56. doi:10.1080/03610927308827006

135. Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In C. E. Brodley, & A. Danyluk (Eds.), *Proceedings of the Eighteenth International Conference on Machine Learning, ICML 2001* (pp. 282-289). San Francisco: Morgan Kaufmann Publishers.

136. Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *Journal of the American Statistical Association, 73*(362), 253-263.

137. Lasswell, H. D., & Kaplan, A. (1950). *Power and society.* New Haven: Yale University Press.

138. Laumann, E. O., Marsden, P. V., & Prensky, D. (1983). The boundary specification problem in network analysis. In R. S. Burt, & M. J. Minor (Eds.), *Applied Network Analysis* (18-34). London: Sage Publications.

139. Lauri, J. (1987). Graph reconstruction - Some techniques and new problems. *Ars Combinatoria, 24*(Serial B), 35-61.

140. Lauri, J. (1992). Vertex-deleted and edge-deleted subgraphs. In R. Ellul-Micallef, & S.Fiorini (Eds.), *Collected papers.* Malta. Retrieved from http://staff.um.edu.mt/jlau/research/UoM400.pdf

141. Lauri, J. (2004). The reconstruction problem. In J.Gross, & J. Yellen (Eds.), *Handbook of graph theory* (pp.79-98). Boca Raton, FL: CRC Press.

142. Law, A. M., & Kelton, W. D. (2000). *Simulation modeling and analysis* (3rd ed.). Boston: McGraw-Hill.

143. Lawrence, P. R., & Lorsch, J. W. (1967). New management job: The integrator. *Harvard Business Review*(November - December 1967), 142-151.

144. Lazarsfeld, P. F. & Henry, N. W. (1968). *Latent structure analysis*, Boston: Houghton Mifflin Company.

145. Lemmer, J. F.(1996). The causal Markov condition, fact or artifact? *SIGART Bulletin, 7*(3), 3-16.

146. Lenski, G. E. (1954). Status crystallization: A non-vertical dimension of social status. *American Sociological Review, 19*, 405-413.

147. Lewin, K. (1952). Group decision and social change. In G. Swanson, T. Newcomb, & E. Hartley (Eds.), *Readings in social psychology* (Rev. ed.) (pp. 459-473) New York: Holt.

148. Liben-Nowell, D. (2005). *An algorithmic approach to social networks.* (Doctoral dissertation, Massachusetts Institute of Technology, 2005). Retrieved from www.cs.carleton.edu/faculty/dlibenno/papers/thesis/thesis.pdf

149. Liben-Nowell, D., & Kleinberg, J. (2004). The link prediction problem for social networks. An updated version of the paper in the *Proceedings of the Twelfth Annual ACM International Conference on Information and Knowledge Management (CIKM'03)* (pp. 556-559). Retrieved February 24, 2007 from http://cs.carleton.edu/faculty/dlibenno/papers/link-prediction/link.pdf

150. Likert, R. (1961). New patterns of management. New York: McGraw-Hill.

151. Little, R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data. New York: Wiley.

152. Liu, W. T., & Duff, R. W. (1972). The strength in weak ties. *Public Opinion Quarterly, 36*(3), 361-366.

153. Long, J. S. (1983). *Confirmatory factor analysis. A preface to LISREL.* Sage university paper series on quantitative applications in the social sciences, No. 07-033. Beverly Hills: Sage Publications.

154. Lowrance, W. W. (1971). *Of acceptable risk.* Los Altos, CA: William Kaufmann.

155. MacDonald, D. (1976). Communication roles and communication networks in a formal organization. *Human Communication Research, 2*(4), 365-375.

156. MacDonald, J. H. (2008). *Handbook of biological statistics.* Baltimore, MD: Sparky House Publishing. Retrieved December 4, 2008, from http://udel.edu/m̃cdonald/statintro.html

157. Magdon-Ismail, M., Goldberg, M., Wallace, W., & Siebecker, D. (2003). Locating hidden groups in communication networks using hidden markov models. In H. Chen, R. Miranda, D. D. Zeng, C. C. Demchak, J. Schroeder, & T. Madhusudan (Eds.) *Proceedings of the first NSF/NIJ symposium on intelligence and security informatics (ISI03)* In *Lecture notes in Computer Science, 2665* (pp. 126-137). New York: Springer.

158. March, J. G. (1953-54). Husband-Wife interaction over political issues. *Public Opinion Quarterly, 17*(Winter Issue), 461-470.

159. March, J. G. (1955). An introduction to the theory and measurement of influence. *The American Political Science Review, 49*(2), 431-451.

160. Manvel, B. (1969). On reconstruction of graphs. In G. T. Chartrand & S. F. Kappor (Eds.), *The Many Facets of Graph Theory* (pp. 207-214). Berlin: Springer.

161. Marsden P. V., & Friedkin, N. E. (1994). Network studies of social influence. In J. Galaskiewicz, & S. Wasserman, *Advances in social and behavioral science from social network analysis* (pp. 3-25). Newbury Park, CA: Sage Publications.

162. Martin J. D., & VanLehn, K. (1994). *Discrete factor analysis: Learning hidden variables in Bayesian networks.* (Technical Report LRGC-ONR-94-1, 1994). Pittsburgh, PA: LRDC, University of Pittsburgh.

163. McCormick, G. H., & Owen, G. (2000). Security and coordination in a clandestine organization. *Mathematical and Computer Modelling, 31*, 175-192.

164. McMullen, B. (2005). *Graph reconstruction numbers.* Master's project report, Rochester Institute of Technology, Rochester, New York. Retrieved November 13, 2007, from https://ritdml.rit.edu/dspace/bitstream/1850/2773/2/ BMcMullenThesis2004.pdf

165. Meek, C. (1995). Causal inference and causal explanation with background knowledge. In P. Besnard, & S. Hanks (Eds.), *Proceedings of the 11th Annual Conference on Uncertainty in AI (UAI-95)* (pp. 403-410). San Mateo, CA: Morgan Kauffman. Retrieved December 5, 2008, from http://ftp.andrew.cmu.edu/pub/phil/chris/causal.ps

166. Merton, R. K. (1949). Patterns of influence: A study of interpersonal influence and of communications behavior in a local community. In P. F. Lazarsfeld & F. N. Stanton (Eds.), *Communications Research, 1948-1949* (pp. 180-219). New York: Harper and Brothers.

167. Merton, R. K. (1957). *Social theory and social structure* (rev. ed.). Glencoe, IL: Free Press.

168. Michalewicz, Z., & Fogel, D. B. (2002). *How to solve it: Modern heuristics.* New York: Springer-Verlag.

169. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2001). *Introduction to linear regression analysis* (3rd. ed.). New York: John Wiley and Sons, Inc.

170. Murphy, K. (1998). *A brief introduction to graphical models and bayesian networks.* Retrieved March 12, 2007 from http://www.cs.ubc.ca/m̃urphyk/Bayes/bnintro.html

171. Murphy, K., & Mian, S. (1999). *Modeling gene expression data using dynamic Bayesian networks.* (Technical Report, 1999) Berkley, CA: Computer Science Division, University of California.

172. Myers, R. H. & D. C. Montgomery (2002). *Response surface methodology: Process and product optimization using designed experiments.* New York: John Wiley and Sons, Inc.

173. Myrvold, W. (1988). Ally and adversary reconstruction problems (Doctoral dissertation, University of Waterloo, Canada, 1988). Abstract retrieved from Dissertation Abstracts Online via FirstSearch.

174. Myrvold, W. (1990). The ally reconstruction number of a tree with five or more vertices is three. *Journal of Graph Theory, 14*(2), 149-166.

175. Myrvold, W. (1992). The degree sequence is reconstructible from n-1 cards. *Discrete Mathematics, 102*, 187-196.

176. Nash-Williams, C. S. J. A. (1978). The reconstruction problem. In L. W. Beineke & R. Wilson (Eds.), *Selected topics in graph theory* (pp. 205236). New York: Academic Press.

177. National Research Council: Committee on Network Science for Future Army Applications (2005). *Network Science.* Washington, DC: The National Academies Press.

178. Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear regression models* (3rd ed.). Chicago: Irwin.

179. Newman, M. E. J. (2006). *Processes taking place on networks* [PDF document]. Retrieved from http://vw.indiana.edu/netsci06/ws-slides/mark_newman.pdf

180. Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence, 29*, 241-288.

181. Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Mateo, CA: Morgan Kauffmann Publishers, Inc.

182. Pearl, J. (2001). *Causality: Models, reasoning and inference.* New York: Cambridge University Press.

183. Pearl, J., Geiger D., & Verma, T. S. (1990). The logic of influence diagrams. In R. M. Oliver & J. Q. Smith (Eds.), *Influence diagrams, belief nets and decision analysis* (pp. 67-88). Rexdale, Ont: Wiley.

184. Pearl, J., & Paz, A. (1987). Graphoids: Graph-based logic for reasoning about relevance relations. In B. Du Boulay, D. Hogg, & L. Steels (Eds.), *Advances in Artificial Intelligence - II* (pp. 357-363). Amsterdam: Elsevier Science Publishers B. V. (North-Holland).

185. Pearl, J., & Verma, T. S. (1991). *A theory of inferred causation.* (Technical Report R-156). Los Angeles: Cognitive Systems Laboratory, Computer Science Department, University of California.

186. Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 49*(12), 1373-1379. doi:10.1016/S0895-4356(96)00236-3

187. Pohar, M., Blas, M., & Turk, S. (2004). Comparison of logistic regression and linear discriminant analysis: A simulation study. *Metodološki zveszki, 1*(1), 143-161. doi:10.1016/S0895-4356(96)00236-3

188. Popescul, A., & Ungar, L. H. (2003). Structural logistic regression for link analysis. In S. Dzeroski, L. De Raedt, & S. Wrobel (Eds.), *Proceedings of the 2nd International Workshop on Multi-Relational Data Mining (MRDM-2003)* (pp. 92-106). New York: ACM Press.

189. Proctor, C. H., & Loomis, C. P. (1951). Analysis of sociometric data. In M. Jahoda, M. Deutsch, & S. Cook (Eds.), *Research methods in social relations* (pp. 561-588). New York: Dryden Press.

190. Project Vote Smart (2008). *Project Vote Smart. The Voter's self-defense system: Senator Harry M. Reid (NV).* Retrieved December 5, 2008, from http://www.votesmart.org/npat.php?can_id=53320

191. Renfro, R. S. II (2001). *Modeling and analysis Of social networks* (Doctoral dissertation, Air Force Institute of Technology, 2001). Retrieved November 17, 2008 from http://www.dtic.mil

192. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work* (pp. 175-186). New York: ACM Press.

193. Richardson, T. (1996). A discovery algorithm for directed cyclic graphs. In E. Horvitz, & F. Jensen (Eds.), *Proceedings of the 12th Annual Conference on Uncertainty in AI (UAI)* (pp. 454-461). San Francisco: Morgan Kauffman.

194. Robins, G., Pattison, P. & Woolcock, J. (2004). Missing data in networks: Exponential random graph (p*) models for networks with non-respondents *Social Networks, 26*, 257-283. doi:10.1016/j.socnet.2004.05.001

195. Robinson, R.W. (1977). Counting unlabled acyclic digraphs. In C.H.C. Little (Ed.), *Lecture notes in mathematics, 622: Combinatorial mathematics V.* New York: Springer-Verlag.

196. Rogers, E. M., & Bhowmik, D. K. (1971). Homophily-Heterophily: Relational concepts for communication research. *Public Opinion Quarterly, 34*, 523-538.

197. Rogers, E. M., & Kincaid, D. L. (1981). *Communication networks. Toward a new paradigm for research.* New York: The Free Press.

198. Rogers, E. M., & Shoemaker F. F. (1970). *Communication of innovations: A cross-cultural approach.* New York: The Free Press.

199. Rogers, R. C., & Maranto, C. L. (1989). Causal models of publishing productivity in psychology. *Journal of Applied Psychology, 74*(4), 636-649.

200. Ross, I. C., & Harary, F. (1955). Identification of the liaison persons of an organization using the structure matrix. *Management Science, 1*(3-4), 251-258.

201. Sageman, M. (2004a). *Global Salafi movement personnel* [Data file].

202. Sageman, M. (2004b). *Understanding terror networks.* Philadelphia, PA: University of Pennsylvania Press.

203. Salmon, W. (1984). *Scientific explanation and the causal structure of the world.* Princeton, NJ: Princeton University Press.

204. Schwartz, D. F., & Jacobson, E. (1977). Organizational communication network analysis: The liaison communication role. *Organizational Behavior and Human Performance, 18*(1), 158-174.

205. Scheines, R., Spirtes, P., Glymour, C., Meek, C., & Richardson, T. (1995). *TETRAD 3: Tools for causal modeling, user's manual.* Retrieved from http://tweedle-dum.phil.cmu.edu/projects/tetrad/tet3/master.htm

206. Scheines, R., Spirtes, P., Glymour, C., Meek, C., & Richardson, T. (1998). The TETRAD project: Constraint based aids to causal model specification. *Multivariate Behavioral Research, 33*(1), 65-117.

207. Seidman, S. B. & Foster, B. L. (1978). A graph-theoretic generalization of the clique concept. *Journal of Mathematical Sociology, 6*, 139-154.

208. Senate of the United States. (2007a). *Committee and Subcommittee Assignments, April 2, 2007* (Senate Publication No. 110-5). Washington, DC: U.S. Government Printing Office.

209. Senate of the United States. (2007b). *Committee and Subcommittee Assignments, December 1, 2007* (Senate Publication No. 110-16). Washington, DC: U.S. Government Printing Office.

210. Shipley, B. (1997). Exploratory analysis with applications in ecology and evolution. *American Naturalist, 149*, 1113-1138.

211. Shipley, B. (2000). *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference.* Cambridge: Cambridge University Press.

212. Shipley, B. (2002). *Cause and correlation in biology: A user's guide to path analysis, structural equations and causal inference.* (Paperback ed.). Cambridge: Cambridge University Press.

213. Silva, R. (2005). *Automatic discovery of latent variable models* (Doctoral dissertation, Carnegie Mellon University, 2005). Retrieved June 6, 2006 from http://www.gatsby.ucl.ac.uk/r̃bas/thesis.pdf

214. Silva, R., Scheines, R., Glymour, C., & Spirtes, P. (2006). Learning the structure of linear latent variable models. *Journal of Machine Learning Research, 7*, 191-246.

215. Simon, H.A. (1952). On the definition of the causal relation. *Journal of Philosophy, 49*, 517-528.

216. Simon, H. A. (1953). Notes on the observation and measurement of political power. *The Journal of Politics, 15*(4), 500-516.

217. Singh, B. K. (1975). Path analysis in social science research. *Indian Journal of Extension Education, 11*, 54-63.

218. Sober, E. (1988). The principle of common cause. In J. H. Fetzer (Ed.) *Probability and causality* (pp. 211-228). Dordecht: D. Reidel.

219. Spirtes, P., Glymour, C., & Scheines, R. (1990). Causality from probability. In J. E. Tiles, G. T. McKee, & G. C. Dean (Eds.), *Evolving knowledge in natural science and artificial intelligence* (pp. 181-199). London: Pitman.

220. Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction and search*. Cambridge, MA: MIT Press.

221. Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction and search* (2nd ed.). Cambridge, MA: MIT Press.

222. Steinley, D., & Wasserman, S. (2006). *Approximate distributions of several common graph statistics: Hypothesis testing applied to a terrorist network.* (Technical Report 06-03, Dec 21, 2006) Bloomington, IN: Department of Statistics, Indiana University.

223. Stockmeyer, P.K. (1977). The falsity of the reconstruction conjecture for tournaments. *Journal of Graph Theory, 1*, 19-25.

224. Stockmeyer, P.K. (1981). A census of non-reconstructable digraphs, I: Six related families. *Journal of Combinatorial Theory B, 31*, 232-239.

225. Stork, D., & Richards, W. (1992). Nonrespondents in communication network studies. *Group & Organization Management, 17*(2), 193-209.

226. Syropoulos, A. Mathematics of multisets. In C. S. Calude, M. J. Dinneen, & G. Păun (Eds.), *Pre-proceedings of the Workshop on Multiset Processing* (pp. 286 - 295). (Centre for Discrete Mathematics and Theoretical Computer Science Research Report 140, August 2000) Auckland, NZ: Department of Computer Science, University of Auckland. Retrieved November, 17, 2008, from http://www.cs.auckland.ac.nz/CDMTCS/researchreports/140WMP00.pdf#page=292

227. Taskar, B., Abbeel, P. & Koller, D. (2002). Discriminative probabilistic models for relational data. In A. Darwiche, & N. Friedman (Eds.), *Proceedings of the 18th Annual Conference on Uncertainty in AI (UAI)* (pp. 485-492). San Francisco: Morgan Kauffman.

228. Taskar, B., Abbeel, P., Wong, M. & Koller, D. (2003). Label and link prediction in relational data. In L. Getoor & D. Jensen (Eds.), *Working Notes of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data (SRL-2003)* (pp. 145-152). Retrieved December 11, 2006 from kdl.cs.umass.edu/srl2003_upload/files/taskar-paper.pdf

229. Taskar, B., Wong, M., & Abbeel, P. & Koller, D. (2004). Link prediction in relational data. In S. Thrun, L. K. Saul, & B. Scholkopf. (Eds.), *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference* (pp. 659-666). Cambridge, MA: MIT Press.

230. Ulam, S.M. (1952). Random processes and transformations. *Proceedings of the International Congress of Mathematicians, 2*, (pp. 264-275). Providence, RI: American Mathematical Society.

231. Ulam, S. (1960). *A collection of mathematical problems.* New York: Wiley.

232. United States Senate (2007). *U.S. Senate roll call votes 110th Congress - 1st session (2007).* Retrieved from http://www.senate.gov/legislative/LIS/roll_call_lists/vote_menu_110_1.htm

233. United States Senate (2008a). *U.S. Senate roll call votes 110th Congress - 2nd session (2008).* Retrieved from http://www.senate.gov/legislative/LIS/roll_call_lists/vote_menu_110_2.htm

234. United States Senate (2008b). *Senate Organization Chart for the 110th Congress.* Retrieved December 5, 2008, from http://www.senate.gov/pagelayout/reference/e_one_section_no_teasers/org_chart.htm

235. United States Senate (2008c). *Freshman senators in the 110th Congress.* Retrieved December 5, 2008, from http://www.senate.gov/galleries/daily/freshmen2.htm

236. United States Senate (2008d). *Senators of the 110th Congress* . Retrieved December 5, 2008, from http://www.senate.gov/general/contact_information/senators_cfm.cfm?OrderBy=party&Sort=ASC

237. United States Senate (2008e). *About Harry Reid. Biography.* Retrieved December 5, 2008, from http://reid.senate.gov/about/index.cfm

238. Verma, T. S., & Pearl, J. (1990). Causal networks: Semantics and expressiveness. In L. Kanal, T. Levitt, R. Shachter, & J. F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence 4*, (pp. 69-76). Amsterdam: Elsevier Science Publishers B. V. (North-Holland).

239. Verma, T. S., & Pearl, J. (1991). Equivalence and synthesis of causal models. In P. P. Bonissone, M. Henrion, L. N. Kanal, & J. F. Lemmer (Eds.), *Uncertainty*

in Artificial Intelligence 6, (pp. 255-268). New York: Elsevier Science Publishers B. V.

240. Vingron, M., Stoye, J., & Luz, H. (2002). *Algorithms for phylogenetic reconstructions. Lecture notes and exercises, Winter 2002/2003* [PDF document]. Retrieved from http://lectures.molgen.mpg.de/Algorithmische_Bioinformatik_WS0405/phylogeny_script.pdf

241. Von Neumann, J. (1966). *Theory of self-reproducing automata* (A. W. Burks, Ed.). Urbana, IL: University of Illinois Press,

242. Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications.* New York: Cambridge University Press.

243. Watts, D. J. (2003). *Small worlds: The dynamics of networks between order and randomness.* Princenton, NJ: Princeton University Press.

244. Weiss, R. S., & Jacobson, E. (1955). A method for the analysis of the structure of complex organizations. *American Sociological Review, 20*, 661-668.

245. West, D. B. (2001). *Introduction to graph theory* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

246. White, H. C., Boorman, S. A., & Breiger, R. L. (1976). Social structure from multiple networks. I. Blockmodels of roles and positions. *American Journal of Sociology, 81*(4), 730-779.

247. Wikipedia Multiset (2008). Retrieved November, 2008, from Wikipedia: http://en.wikipedia.org/wiki/Multiset

248. Winston, W. L. (1994) *Operations research: Applications and algorithms* (3rd ed.). Belmont, CA: Wadsworth Publishing Company.

249. Zhang, N. L. (2004). Hierarchical latent class models for cluster analysis, *Journal of Machine Learning Research, 5*, 697-723.

250. Zhang, N. L., Nielsen, T. D., & Jensen, F. V. (2004). Latent variable discovery in classification models. *Artificial Intelligence in Medicine, 30*, 283-299.

251. Zwick, M. (2004). An overview of reconstructability analysis. *Kybernetes, 33*(5/6), 877-905.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of the collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to an penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* 26-03-2009	**2. REPORT TYPE** **Doctoral Dissertation**	**3. DATES COVERED** *(From – To)* Sep 2004 – Jan 2009

4. TITLE AND SUBTITLE

CHARACTERIZING AND DETECTING UNREVEALED ELEMENTS OF NETWORK SYSTEMS

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Leinart, James, A., Lieutenant Colonel, USAF

5d. PROJECT NUMBER
09ENS205

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Street, Building 642
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/DS/ENS/08-01W

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This dissertation addresses the problem of discovering and characterizing unknown elements in network systems. It is not uncommon to have incomplete knowledge of network systems due to either passive circumstances, e.g. limited resources to observe a network, or active circumstances, e.g. intentional acts of concealment, or some combination of active and passive influences. This research suggests statistical and graph theoretic approaches for such situations, including those in which nodes are causally related. A related aspect of this research is accuracy assessment. It is possible an analyst could fail to detect a network element, or be aware of network elements, but incorrectly conclude the associated network system structure. Consequently, this dissertation provides a framework to evaluate accuracy.

15. SUBJECT TERMS
Graphs, Networks, Detection, Causality, Social Sciences, Reconstruction

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Richard F. Deckro, Civ, USAF (ENS)
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-3636, ext 4325; e-mail: Richard.Deckro@afit.edu
U	**U**	**U**	**UU**	158	